dgfs
Deutsche Gesellschaft
für Sprachwissenschaft

TECHNISCHE
UNIVERSITÄT
DARMSTADT

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Tutorium der Sektion CL:

Einführung in die Statistik für Linguisten mit „R“

34. Jahrestagung der DGfS
06. März 2012

Stefan Evert (TU Darmstadt)
Amir Zeldes (HU Berlin)

Worum geht's?

- Wir sind im Alltagsleben und in der Forschung von Statistik umgeben
- Quantitative Aussagen werden oft akzeptiert, ohne dass man sie wirklich versteht
- Möglichkeiten, anhand der Statistik zu neuen Erkenntnissen zu kommen, werden vernachlässigt
- Als Geisteswissenschaftler schwer einzusteigen (aus eigener Erfahrung …)

Wozu? Ein Beispiel

Osnabrück - "Der Trend zur Gewalt ist ungebrochen. Besonders die Zahl gefährlicher und schwerer Körperverletzungen ist deutlich gestiegen", sagte der Bundesvorsitzende der Gewerkschaft der Polizei (GdP), Konrad Freiberg, der "Neuen Osnabrücker Zeitung". Zwar hätten Niedersachsen, Mecklenburg-Vorpommern und das Saarland ihre Kriminalstatistiken noch nicht vorgelegt, die Tendenz für den Bund sei dennoch eindeutig.
http://www.spiegel.de/panorama/justiz/0,1518,473433,00.html

Wozu? Ein Beispiel

Osnabrück - "Der Trend zur Gewalt ist ungebrochen. Besonders die Zahl gefährlicher und schwerer Körperverletzungen ist deutlich gestiegen", sagte der Bundesvorsitzende der Gewerkschaft der Polizei (GdP), Konrad Freiberg, der "Neuen Osnabrücker Zeitung". Zwar hätten Niedersachsen, Mecklenburg-Vorpommern und das Saarland ihre Kriminalstatistiken noch nicht vorgelegt, die Tendenz für den Bund sei dennoch eindeutig.
http://www.spiegel.de/panorama/justiz/0,1518,473433,00.html

Noch ein Beispiel

Die Arzneimittelausgaben der gesetzlichen Krankenkassen steigen im kommenden Jahr voraussichtlich um 6,6 Prozent auf einen Rekordwert von mehr als 31 Milliarden Euro
http://www.sueddeutsche.de/politik/139/313047/text/

- Sie steigen sicher nicht genau um 6.6%.
- Mit welcher Wahrscheinlichkeit ist es 0.5% mehr?
- Mit welcher Wahrscheinlichkeit gibt es keinen Anstieg?

- ➢ So wie es dasteht, kann es nicht eintreffen
- ➢ Keinerlei Überprüfbarkeit

Professionelle Darstellung solcher Zahlen

Die Ergebnisse für die zwei experimentellen Bedingungen wiesen signikante Differenzen auf. Schüler, die nach der neuen Methode unterrichtet wurden, erreichten signifikant bessere Ergebnisse als die nach der traditionellen Methode unterrichteten ($t = 6,03$, $df = 13$, $p < 0.001$)

- Was ist signifikant?
- Was bedeutet t, df, und p?
- Ist das toll?

Zwischenfazit

- Seinem eigenen Gefühl kann man bei der Beurteilung von Zahlenreihen nicht trauen
- Viele empirische Fragen sind ohne Statistik schlicht nicht zu beantworten
- Ernstzunehmende Aussagen über Zahlen sind ohne Ausbildung unverständlich:
  - der eigenen Anschauung nicht zu trauen
  - sich Mittel und Wege anzueignen, in Ihrer eigenen Forschung Zahlen angemessen zu deuten
  - fremde Zahlen zu verstehen und zu beurteilen

Statistik in der Linguistik

- Was ist der Unterschied zwischen gesprochener und geschriebener Sprache?
- unterschiedlichen Textsorten? Geschlecht? …
- Wie ähnlich werden bestimmte Wörter oder Konstruktionen gebraucht? Was ist Gebrauch?
- Wie produktiv sind Wortbildungsmuster im Vergleich? Was ist lexikalisiert?
- Was fällt Deutschlernern besonders schwer?
- Wie kann man Bedeutung empirisch erschließen?
- …

Was ist R?

- Wir werden in diesem Tutorium auch mit echten Daten arbeiten und statistische Tests anwenden
- Der einzig sinnvolle Weg ist, ein professionelles Statistikprogramm zu lernen
- Hier verwenden wir R: http://cran.r-project.org/
- Text- und kommandozeilenbasiert
- Look & Feel unterscheidet sich ziemlich stark von den üblichen GUI-Programmen
- Der Einstieg in die Arbeit mit R erscheint schwieriger

Warum R?

- Praktischer Grund: R ist frei (SPSS bspw. sehr teuer)
- Unheimlich gute graphische Möglichkeiten
- Extrem flexibel, alles mit allem kombinierbar
- Erweiterbar durch tausende von Paketen
- Läuft gleichermaßen unter Windows/Linux/Mac
- Erobert sich in der Wissenschaft mehr und mehr eine beherrschende Stellung
- Für Linguistik besonders beliebt (Module zur Verarbeitung linguistischer Daten)

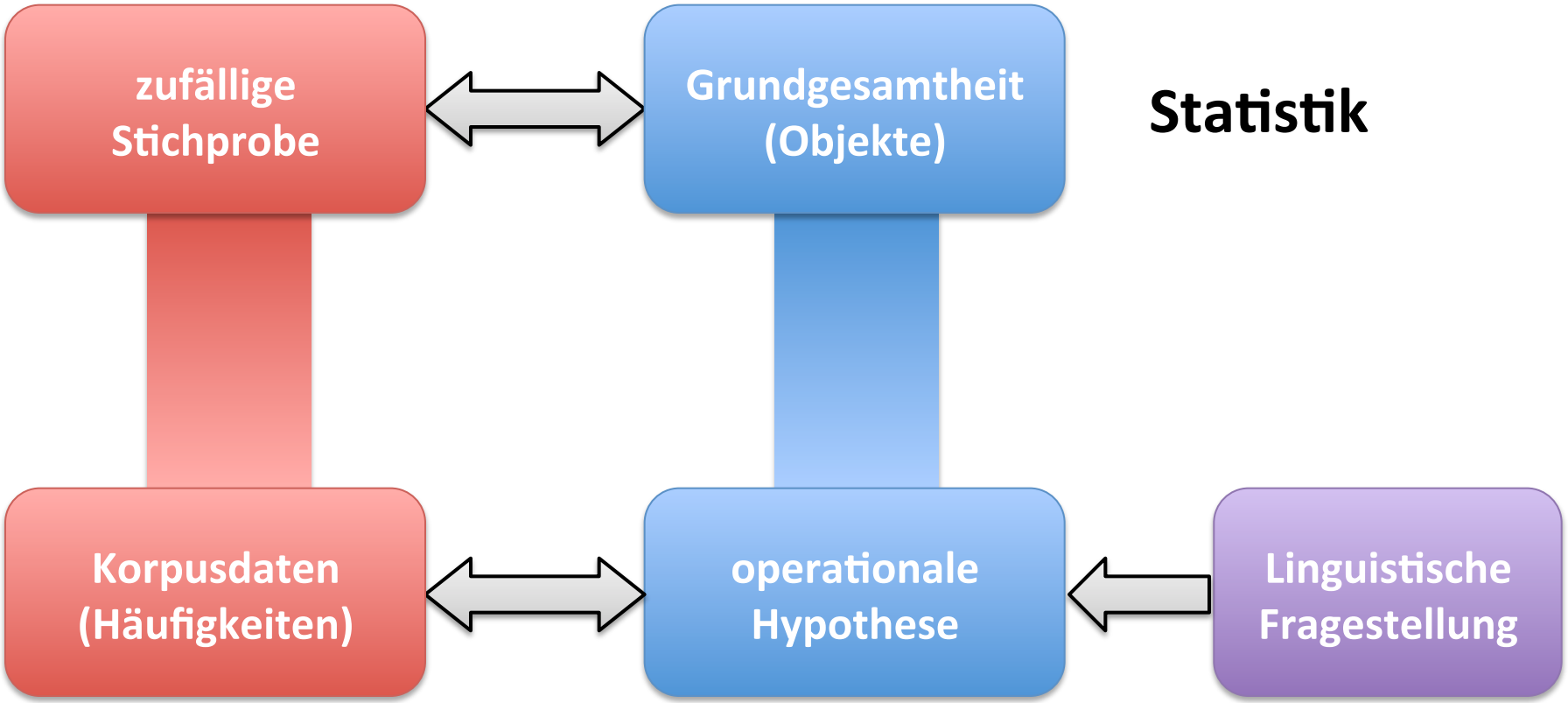
Ablauf des Tutoriums

1. Vergleich von Korpushäufigkeiten
2. Übungen: Nominalkomposita
3. Kaffeepause
4. Kreuztafeln und Assoziation
5. Übungen: Informationsstruktur & Topikalität

- Haben alle Teilnehmer R installiert?

HÄUFIGKEITSVERGLEICH

Quantitative Korpusstudien



Ein Fallbeispiel

Linguistische Fragestellung

- Nominalkomposita bei DaF-Lernern (L2) und deutschen Muttersprachlern (L1)
- Vermutungen
  - L2 bilden weniger Komposita als L1
  - Abhängigkeit von der L1 des Lerners
- Quantitative Studie auf Basis des Lernerkorpus Falko (HU Berlin, Reznicek et al. 2010)

Ein Fallbeispiel

operationale Hypothese

- Operationalisierung:
  Die Häufigkeit von Nominalkomposita ist bei L2-Sprechern geringer als bei L1-Sprechern
- Was bedeutet „Häufigkeit“?
  - Anzahl von Nominalkomposita pro Text?
  - durchschnittliche Anzahl von NK pro Satz?
  - relative Häufigkeit von Nominalkomposita
    = Anteil von NK unter allen Substantiven

Ein Fallbeispiel

operationale Hypothese

- Operationalisierung:
  Die Häufigkeit von Nominalkomposita ist bei L2-Sprechern geringer als bei L1-Sprechern
- Was bedeutet „Häufigkeit“?
  - Anzahl von Nominalkomposita pro Text?
  - durchschnittliche Anzahl von NK pro Satz?
  - relative Häufigkeit von Nominalkomposita
    = Anteil von NK unter allen Substantiven
- In Bezug auf welche Texte? schriftl./mündl.?

zufällige Stichprobe

Ein Fallbeispiel

Korpusdaten (Häufigkeiten)

- Wir benötigen zwei Stichproben von Nomina
  1. aus Texten von deutschen Muttersprachlern
  2. aus Texten von DaF-Lernern
- Was ist eine zufällige Stichprobe?
  - zusammenhängender Text ≠ Zufallsstichprobe
  - Stichproben müssen repräsentativ (für die jeweilige Sprachvarietät) und vergleichbar sein
  - hier: Material aus Lernerkorpus Falko (deutsche Essays von L1 und L2 zu gleichen Themen)

zufällige Stichprobe

Ein Fallbeispiel

Korpusdaten (Häufigkeiten)

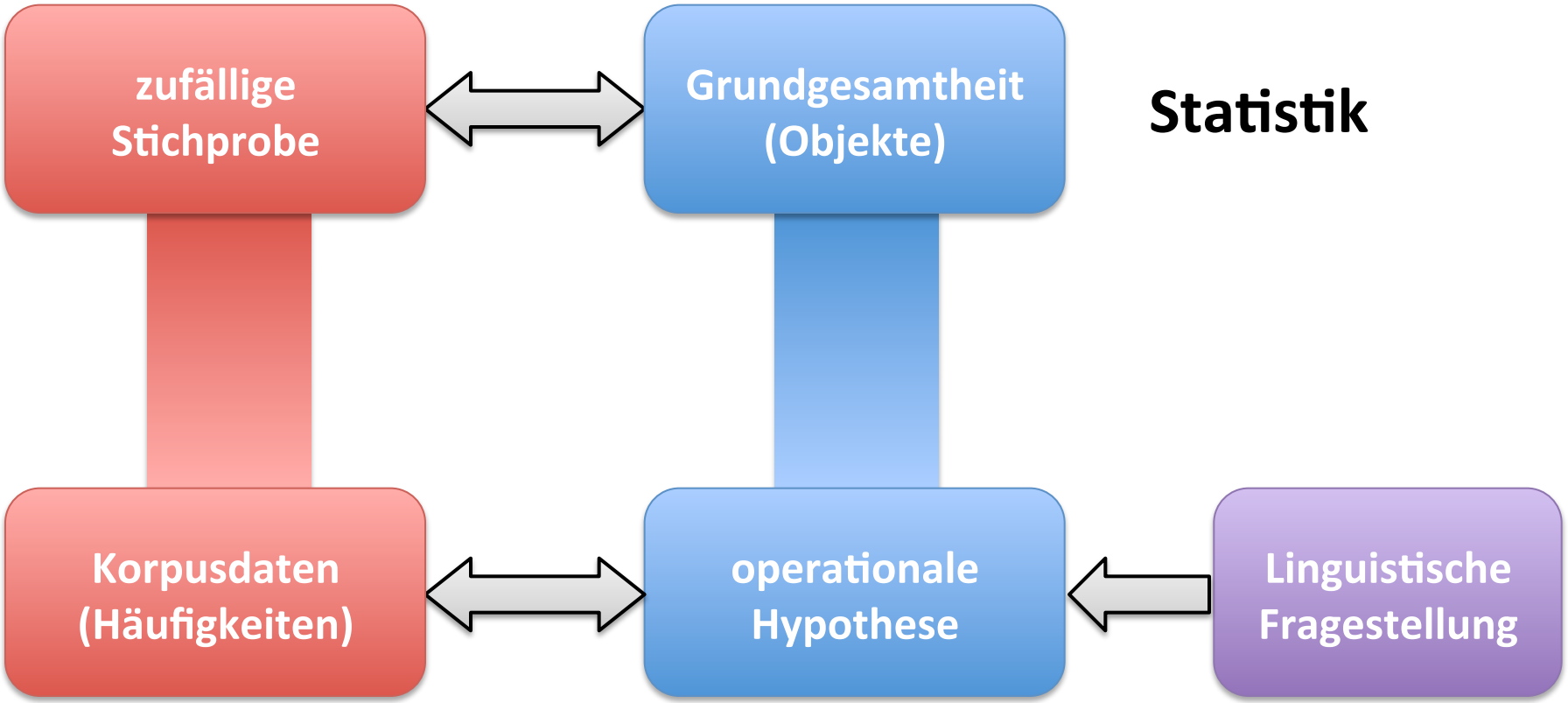
- Stichprobe von unterschiedlichen Wörtern oder einzelnen Vorkommen im Text?
  - Token ⇔ Gebrauchshäufigkeit
    vs. Typen ⇔ Vokabulargröße / Produktivität
  - üblicherweise Stichprobe von Token
  - Produktivitätsmessung erfordert komplexere statistische Methoden (→ Zipfsches Gesetz, LNRE)

Grundgesamtheit?

Grundgesamtheit (Objekte)

- Was genau ist die „Grundgesamtheit“?
  - Wir interessieren uns für Eigenschaften von Sprechern (L1 und L2)
- Extensionaler Sprachbegriff:
  Sprache als Menge von Äußerungen
  - alle tatsächlichen und denkbaren Äußerungen bzw. Texte aus der relevanten Sprachvarietät
  - Objekte = Token (hier: Substantive im Text)
- Statistik trifft Aussagen über Grundgesamtheit!

Korpusstudie: Nominalkomposita



Vereinfachung

- Annahme: Wir kennen bereits die Häufigkeit von Nominalkomposita bei L1
  - aus bereits publizierten Untersuchungen
  - Ergebnis: 16% Nominalkomposita (hypothetisch!)
- Nur noch eine Stichprobe erforderlich
  - Substantive aus Texten von DaF-Lernern
  - wichtig: gleiche Textsorte, Domäne, usw.

Nullhypothese H_0

- Präzise quantitative Formulierung der operationalen Hypothese erforderlich
- Was ist die einfachste mögliche Hypothese?
  - L2 bilden < 16% Nominalkomposita?
  - L2 bilden ≠ 16% Nominalkomposita?
  - L2 bilden auch genau 16% Nominalkomposita?

Nullhypothese H_0

- Präzise quantitative Formulierung der operationalen Hypothese erforderlich
- Was ist die einfachste mögliche Hypothese?
  - L2 bilden < 16% Nominalkomposita
  - L2 bilden ≠ 16% Nominalkomposita
  - L2 bilden auch genau 16% Nominalkomposita
- Nullhypothese H_0 soll widerlegt werden!
  - statistische Verfahren können Hypothesen nur ablehnen, nicht bestätigen

Nullhypothese H_0

- Mathematische Formulierung von H_0:
  Die Häufigkeit μ von Nominalkomposita bei L2 beträgt genau 16%
- In Formeln:

$$H_0 : \ \mu = 16\% = .16$$

Stichprobe

- Zufallstichprobe von $n = 100$ Substantiven
  - Wie erstellt man eine zufällige Stichprobe aus Texten von DaF-Lernern?
  - Erinnerung: Stichprobe muss repräsentativ sein
- Ergebnis: $k = 12$ Komposita
  - Erwartung unter H_0: $k_0 = 16$ Komposita
  - weniger Komposita als erwartet → H_0 widerlegt?
- Zweite Stichprobe: $k = 17$ Komposita
  - Ablehnung von H_0 wäre voreilig gewesen!

Zufallsschwankungen

- Anzahl von Komposita in Stichprobe unterliegt Zufallsschwankungen
  - weicht i.d.R. vom „tatsächlichen“ Wert ab
  - zufällige Auswahl stellt sicher, dass im Mittel die erwartete Anzahl gefunden wird (falls H_0 gilt)

Zufallsschwankungen

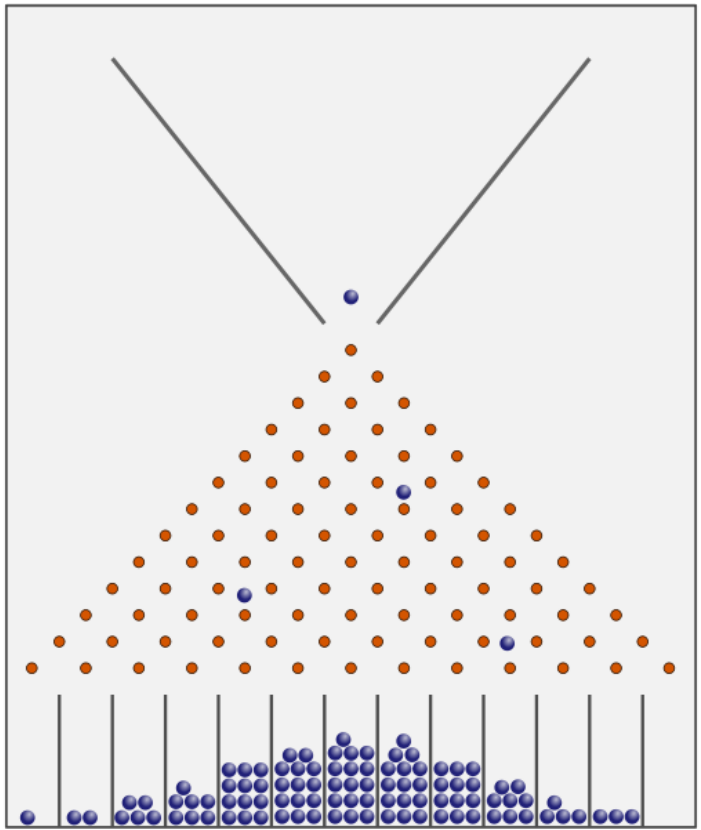
- Bedeutung von $k = 12$ in Stichprobe
  - *a)* H_0 stimmt nicht, tatsächliche Häufigkeit geringer
  - *b)* H_0 stimmt, aber Stichprobe enthält zufällig weniger Komposita als erwartet
- Wir können (a) nur dann folgern, wenn (b) sehr unwahrscheinlich ist
  - intuitiv: Risiko, falsches Ergebnis zu publizieren
  - übliche Kriterien: Risikio < 5%, 1% oder 0.1%

Stichprobenverteilung

- Wie groß ist das Risiko, $k = 12$ oder weniger Komposita zu finden, sofern H_0 stimmt?
  → Stichprobenverteilung (unter H_0)
- Empirisch: viele Stichproben (mit $n = 100$)
  - können als Hausarbeiten vergeben werden ;-)
- Einfacher: Münzwürfe bzw. Würfeln
  - Auswahl von 100 Token = 100-maliges Würfeln
  - jeweils 16% Wahrscheinlichkeit für Kompositum (unter H_0) entspricht ungefähr Wurf von 6 Augen

Stichprobenverteilung

- Simulation von Münzwürfen: Galtonbrett
  - für Japan-Fans: Pachinko-Maschine



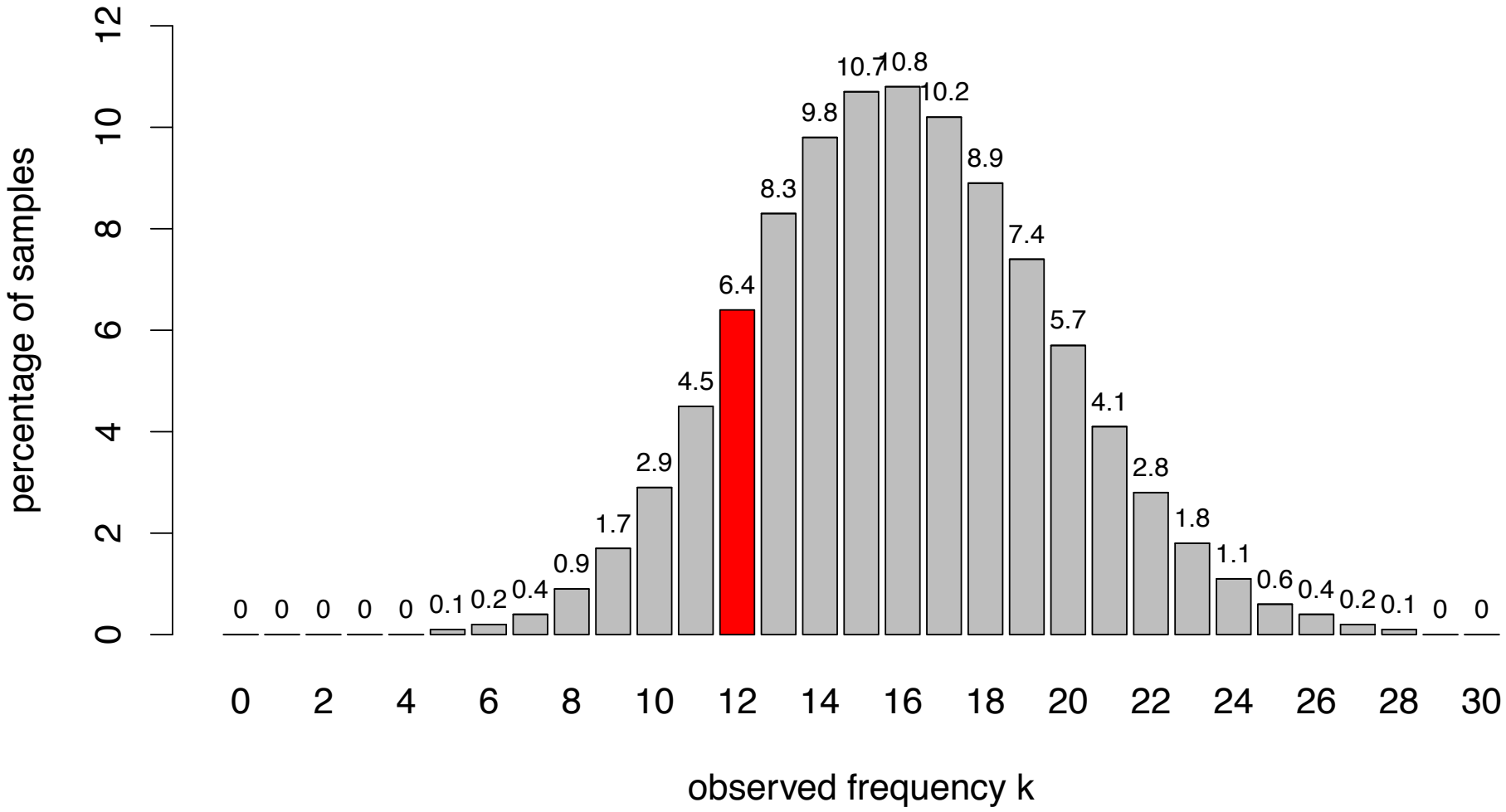
Stichprobenverteilung

- Wir können die Stichprobenverteilung auch ohne Galtonbrett berechnen
- Binomialverteilung B(n, μ)
  - für Stichprobe von n Token
  - bei tatsächlicher Häufigkeit μ (H_0: $\mu = 16\%$)

$$\Pr(k) = \binom{n}{k} \mu^k (1-\mu)^{n-k}$$

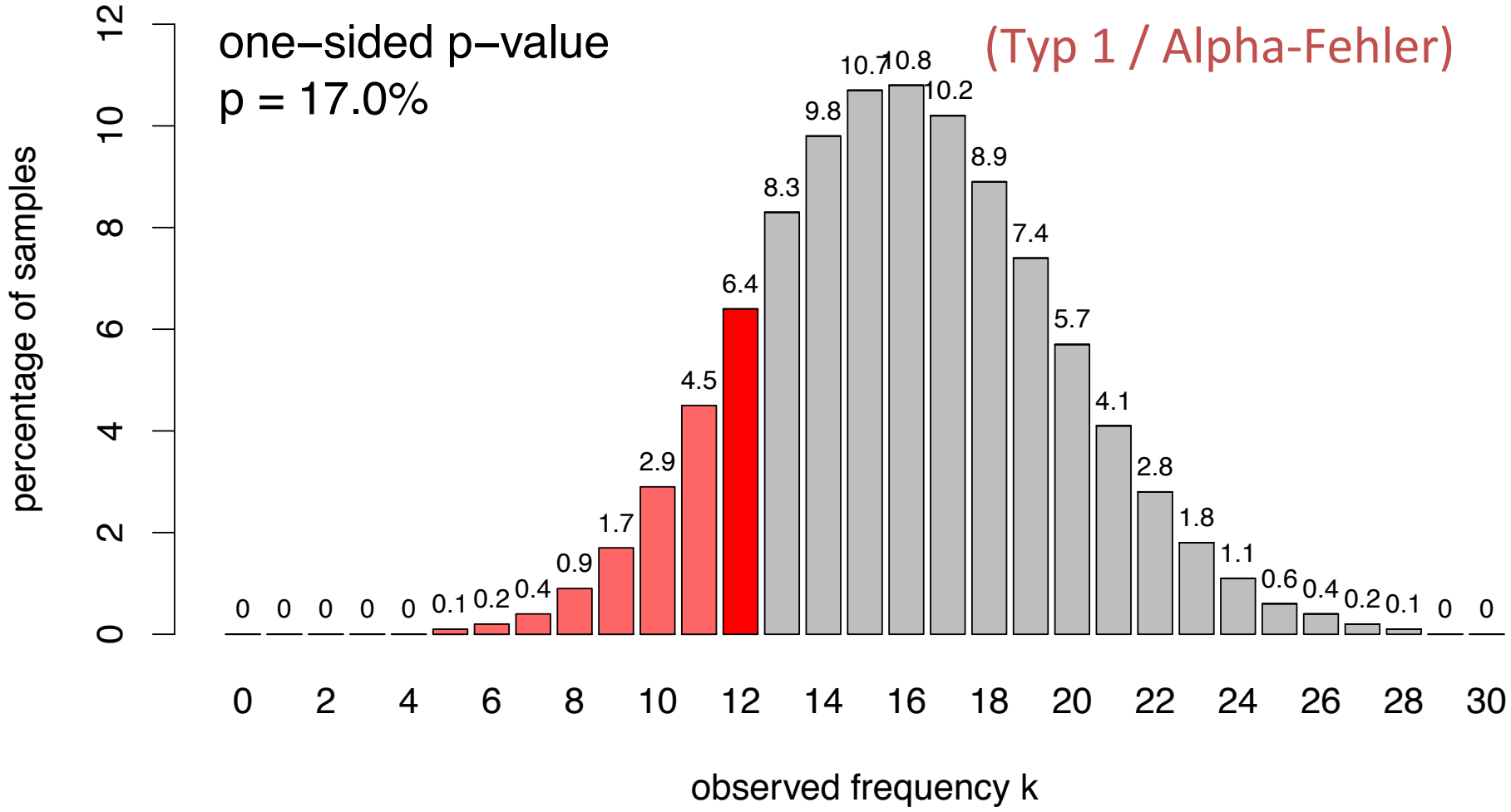
Prozentsatz von Stichproben mit genau k Komposita

Binomialverteilung graphisch



Binomialverteilung graphisch

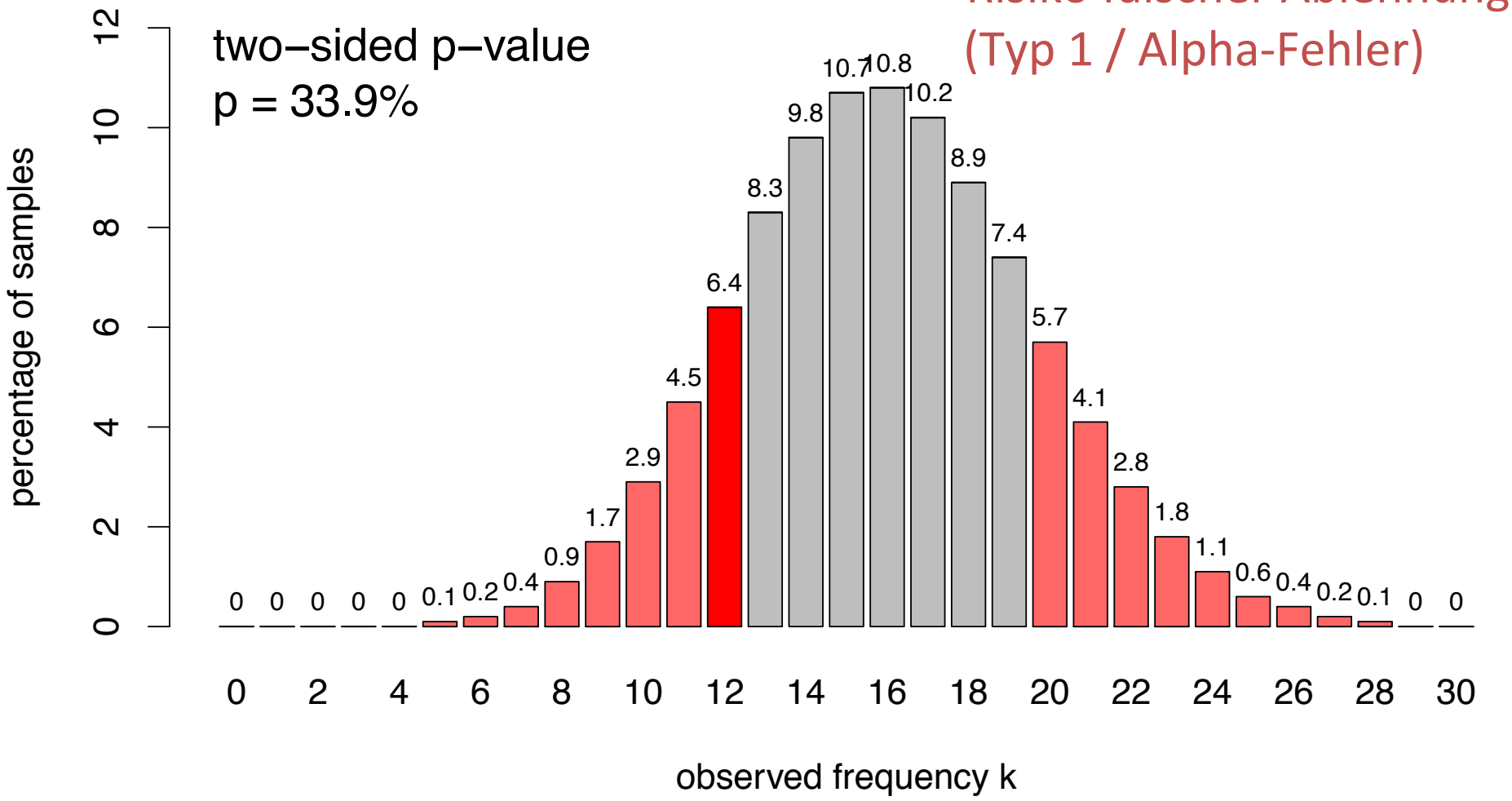
Signifikanzwert = p-value
= Risiko falscher Ablehnung
(Typ 1 / Alpha-Fehler)



Binomialverteilung graphisch

zweiseitiger Test (empfohlen)

Signifikanzwert = p-value
= Risiko falscher Ablehnung
(Typ 1 / Alpha-Fehler)



Signifikanzniveau

- Wann darf H_0 abgelehnt werden?
  - Signifikanzwert *p* beziffert das Fehlerrisiko
  - Ermessensfrage: welches Risiko ist akzeptabel?
- Übliche Signifikanzniveaus
  - $p < .05$ (5%) *
  - $p < .01$ (1%) **
  - $p < .001$ (0.1%) ***

Binomialtest in R

```
> binom.test(12, n = 100, p = .16)
    Exact binomial test

data:  12 and 100
number of successes = 12, number of trials = 100,
    p-value = 0.3392
alternative hypothesis: true probability of success is
    not equal to 0.16
95 percent confidence interval:
 0.0635689 0.2002357
sample estimates:
probability of success
                  0.12
```

Trennschärfe

- Was wäre bei einer 5x größeren Stichprobe?

  $n = 500$, $k = 60$ → relative Häufigkeit $k/n = 12\%$

```
> binom.test(60, n = 500, p = .16)
```

  $p = .01452$* (1.5%)

- Typ II / Beta-Fehler
  - H_0 ist falsch, kann aber nicht abgelehnt werden
  - z.B. tatsächliche Häufigkeit von 12% bei L2
  - größere Stichprobe → bessere Trennschärfe (*power*), d.h. kleinere Abweichung von H_0 (Effektgröße) genügt

Häufigkeitsvergleich

- Wie lautet H_0, wenn Kompositahäufigkeit bei L1 nicht bekannt ist?

$$H_0 : \ \mu_1 = \mu_2$$

- Häufigkeitsvergleich von 2 Stichproben
  - keine Annahme über genauen Wert von $\mu_1 = \mu_2$
  - vergleichbare Stichproben aus L1- und L2-Texten, aber Stichprobengröße darf unterschiedlich sein
- R-Befehl: `prop.test()`

Häufigkeitsvergleich

- Stichprobe L2: $k = 52$, $n = 500$
- Stichprobe L1: $k = 76$, $n = 500$

```
> prop.test(c(52, 76), c(500, 500))
```

Two samples: frequency comparison

	Frequency count	Sample size
Sample 1	52	500
Sample 2	76	500

Clear fields

Calculate

95% confidence interval

in automatic format

with 4 significant digits

Häufigkeitsvergleich

- Stichprobe L2: $k = 52$, $n = 500$
- Stichprobe L1: $k = 76$, $n = 500$

```
> prop.test(c(52, 76), c(500, 500))

    2-sample test for equality of proportions with …

data:  c(52, 76) out of c(500, 500)
X-squared = 4.7395, df = 1, p-value = 0.02948
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.091306444 -0.004693556
sample estimates:
prop 1 prop 2
 0.104  0.152
```

Konfidenzintervall

- Und wenn keine Hypothese vorliegt?
  - „Wie häufig bilden L2-Sprecher Komposita?“
- Häufigkeitsschätzung auf Basis einer Stichprobe von $n = 1000$ Substantiven
  - Ergebnis: $k = 120$, $n = 1000$
  - direkter Schätzwert = Punktschätzer:

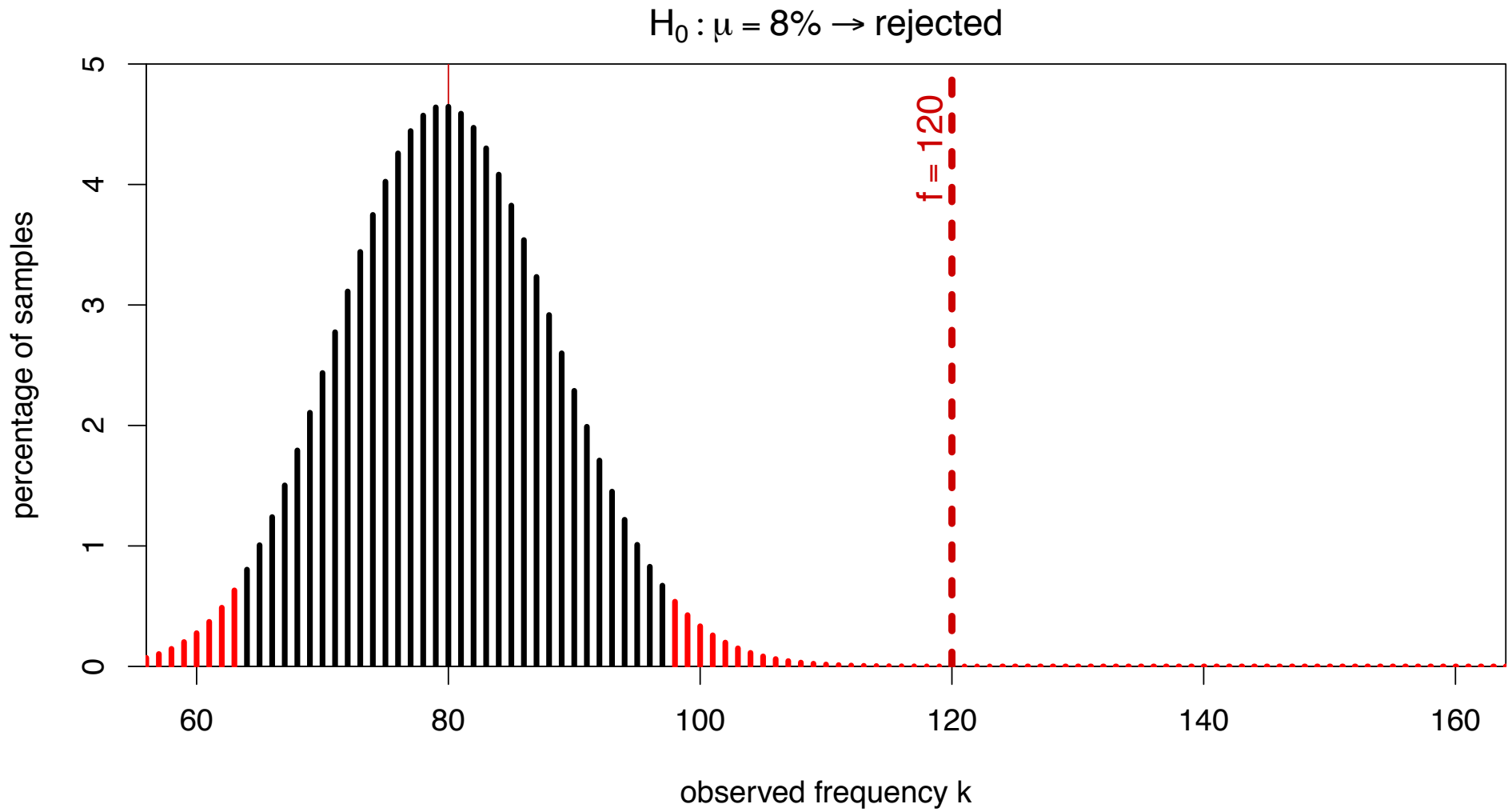
$$\hat{\mu} = \frac{k}{n} = \frac{120}{1000} = .12$$

- Wie zuverlässig ist dieser Schätzwert?

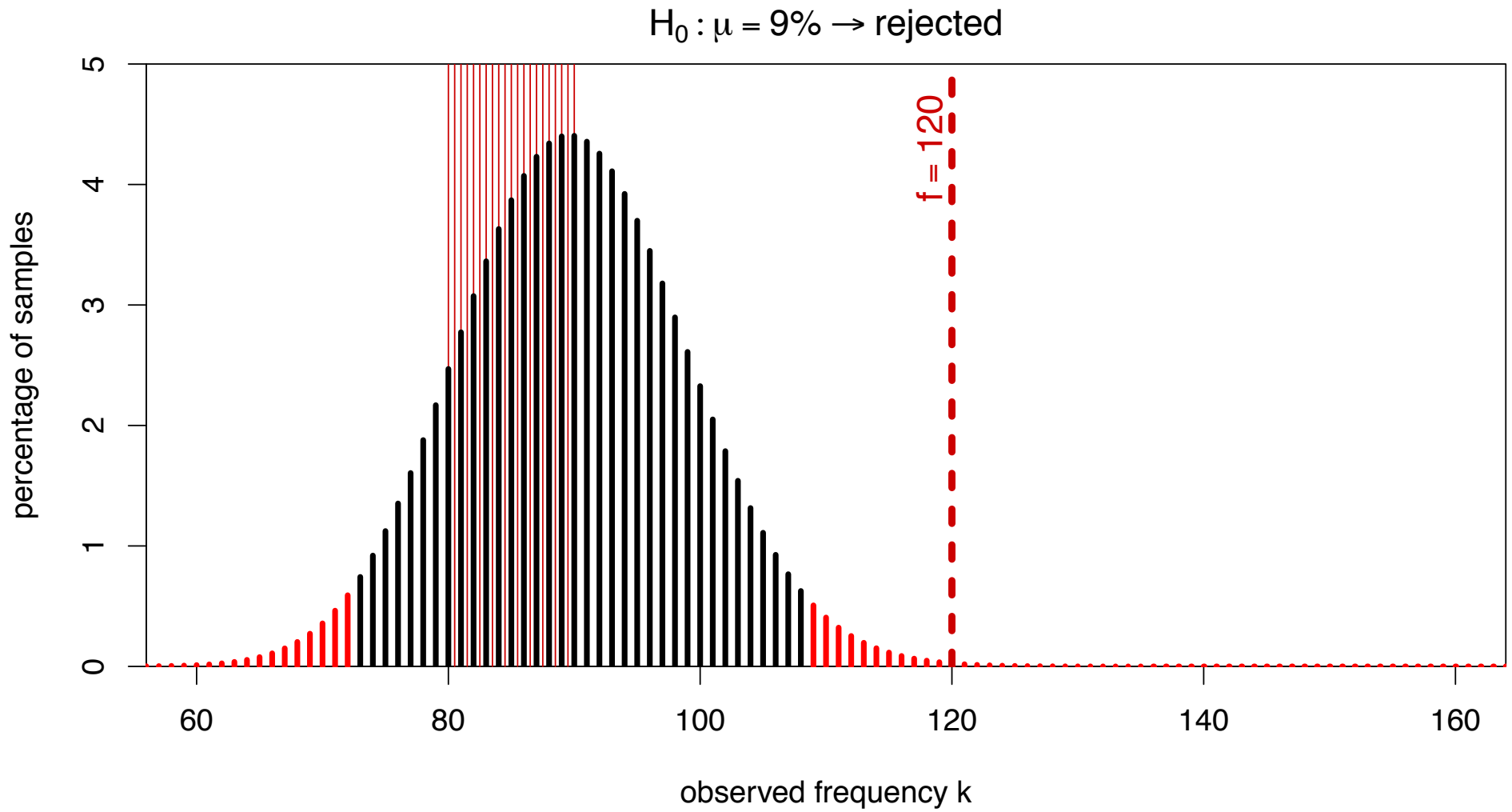
Konfidenzintervall

- Wiederholte Stichproben (jeweils $n = 1000$)
  1) $k = 120 \rightarrow \mu = 12.0\%$
  2) $k = 105 \rightarrow \mu = 10.5\%$
  3) $k = 129 \rightarrow \mu = 12.9\%$
  4) $k = 126 \rightarrow \mu = 12.6\%$
  5) $k = 111 \rightarrow \mu = 11.1\%$
  6) $k = 92 \rightarrow \mu = 9.2\%$
  7) $k = 117 \rightarrow \mu = 11.7\%$

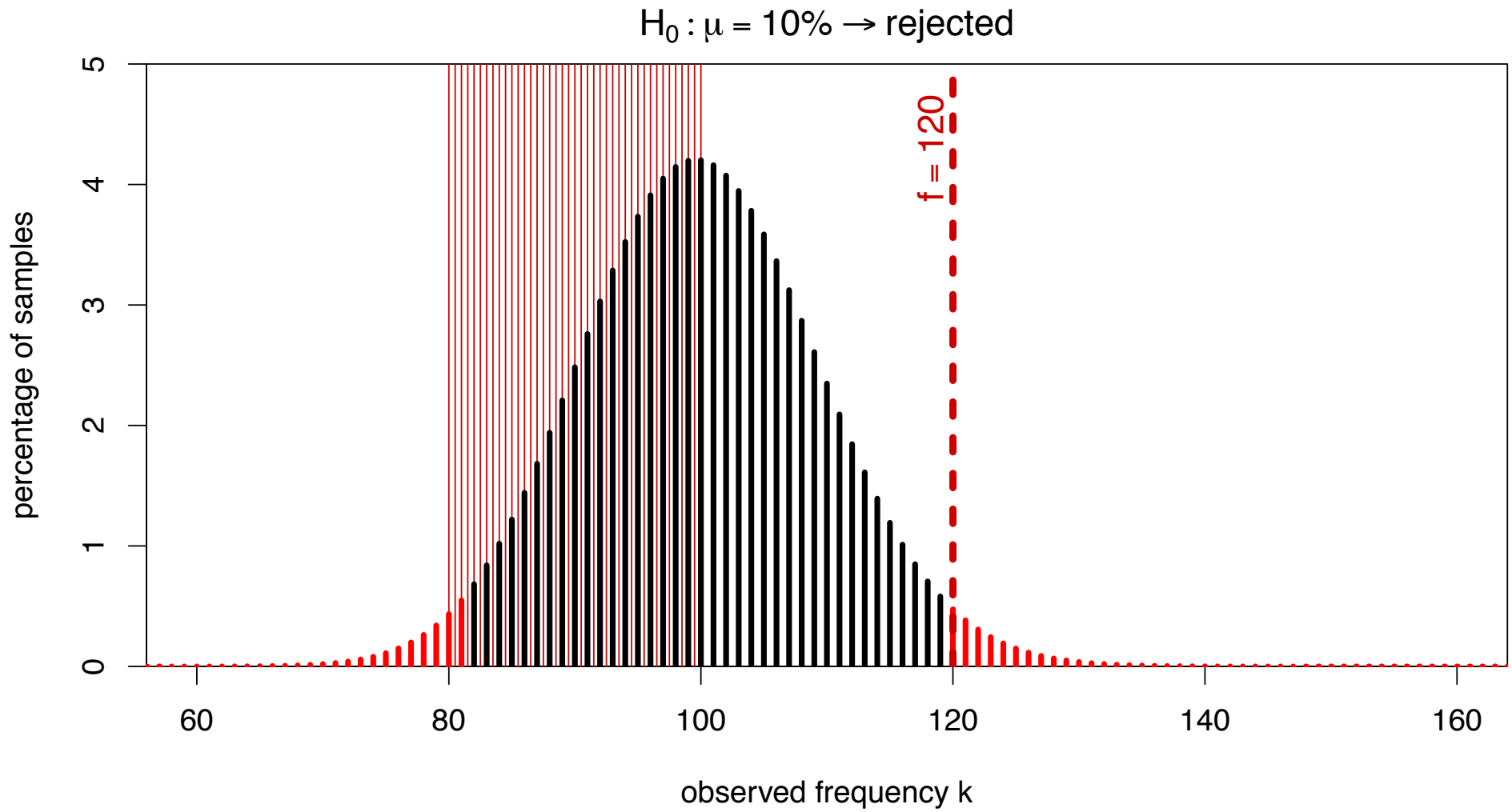
Konfidenz: invertierter Test



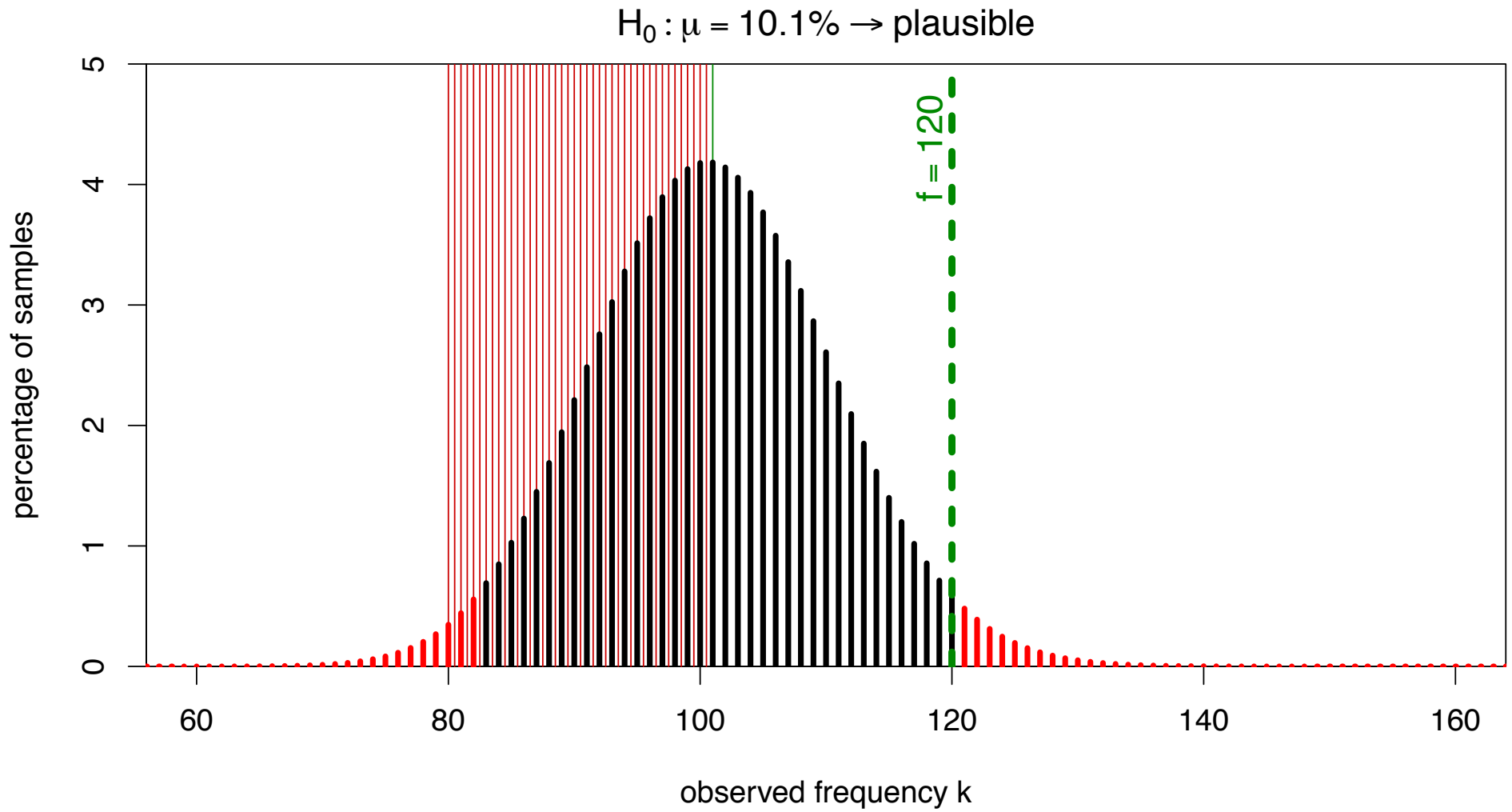
Konfidenz: invertierter Test



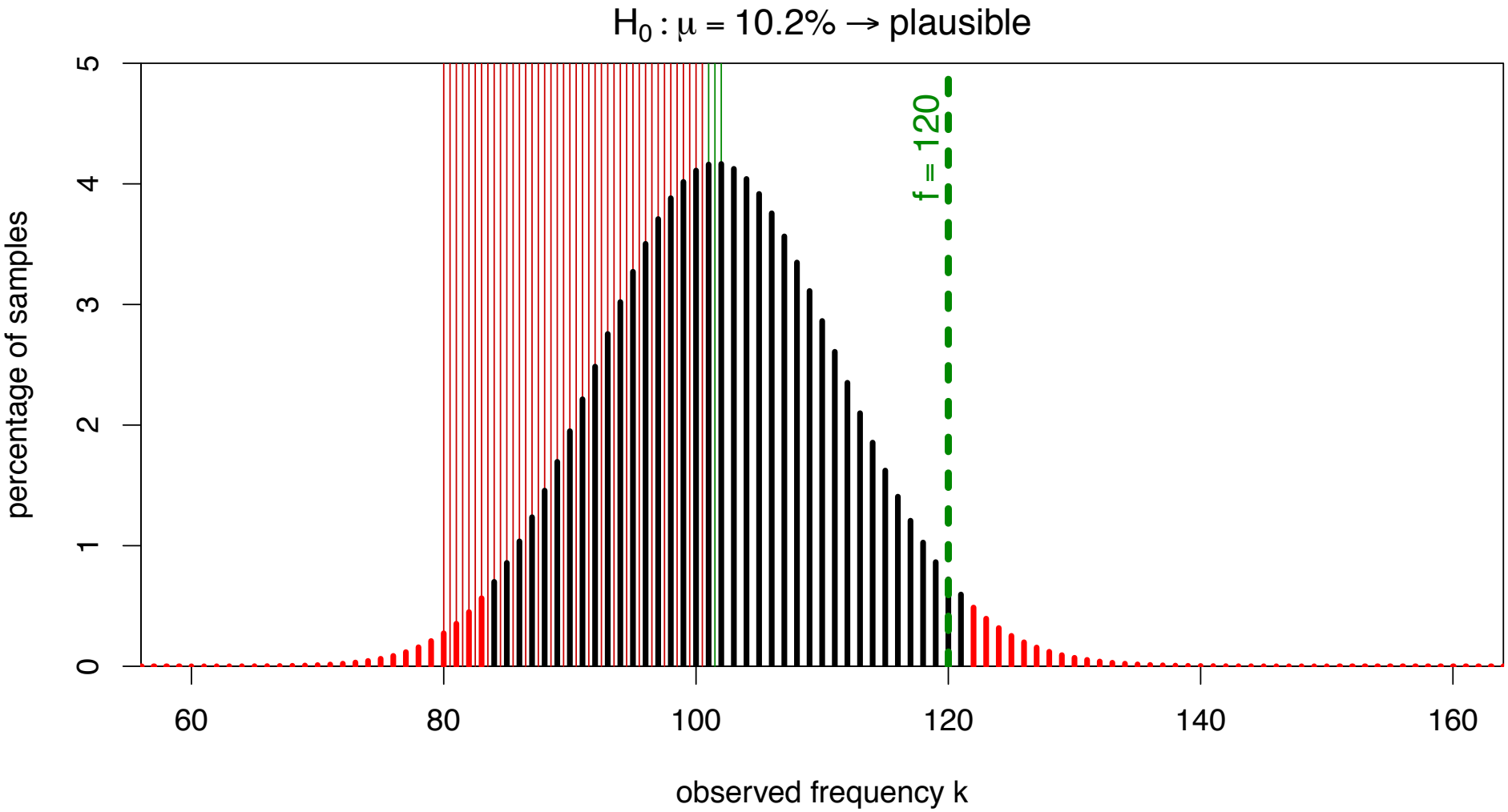
Konfidenz: invertierter Test



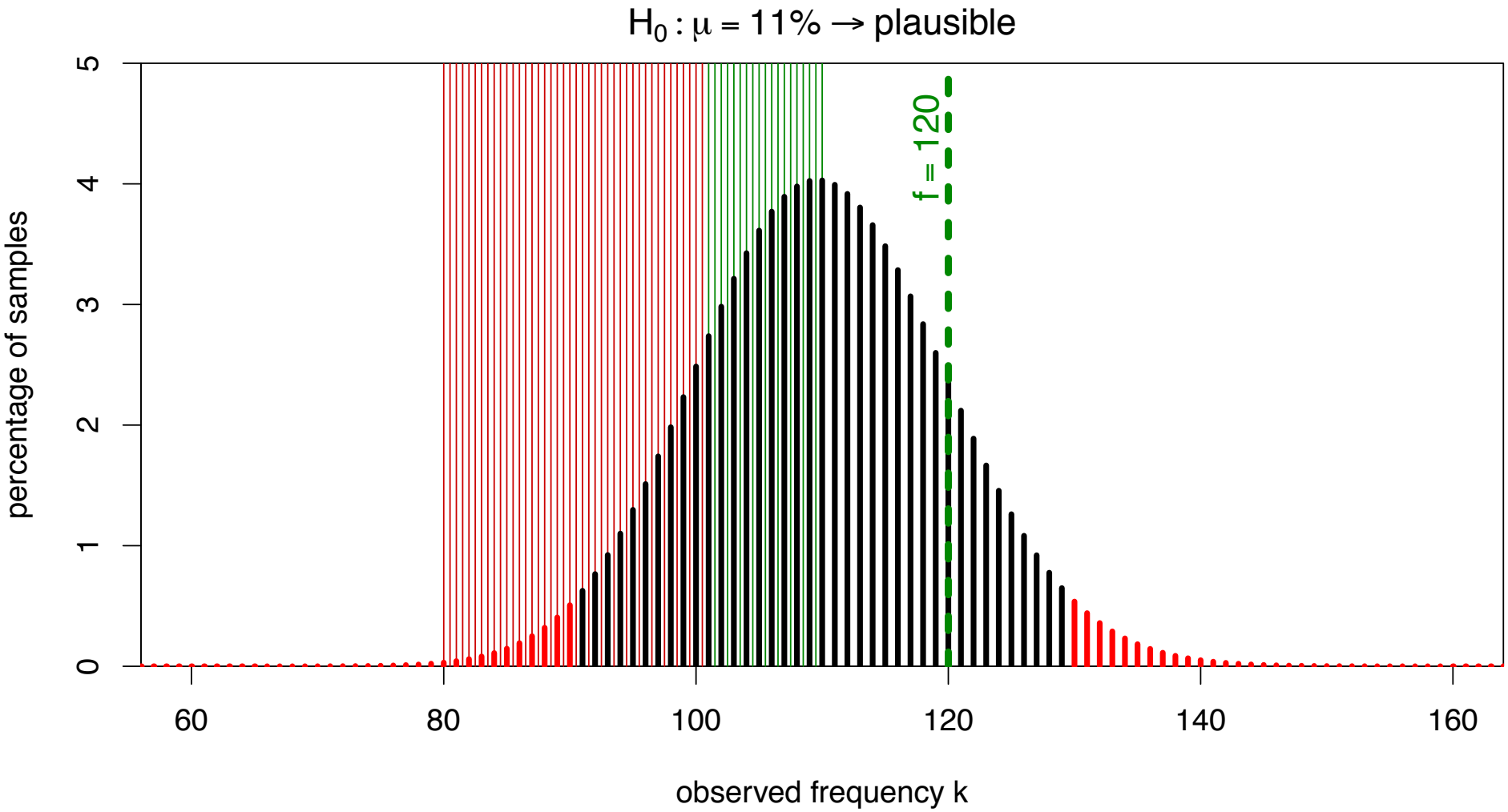
Konfidenz: invertierter Test



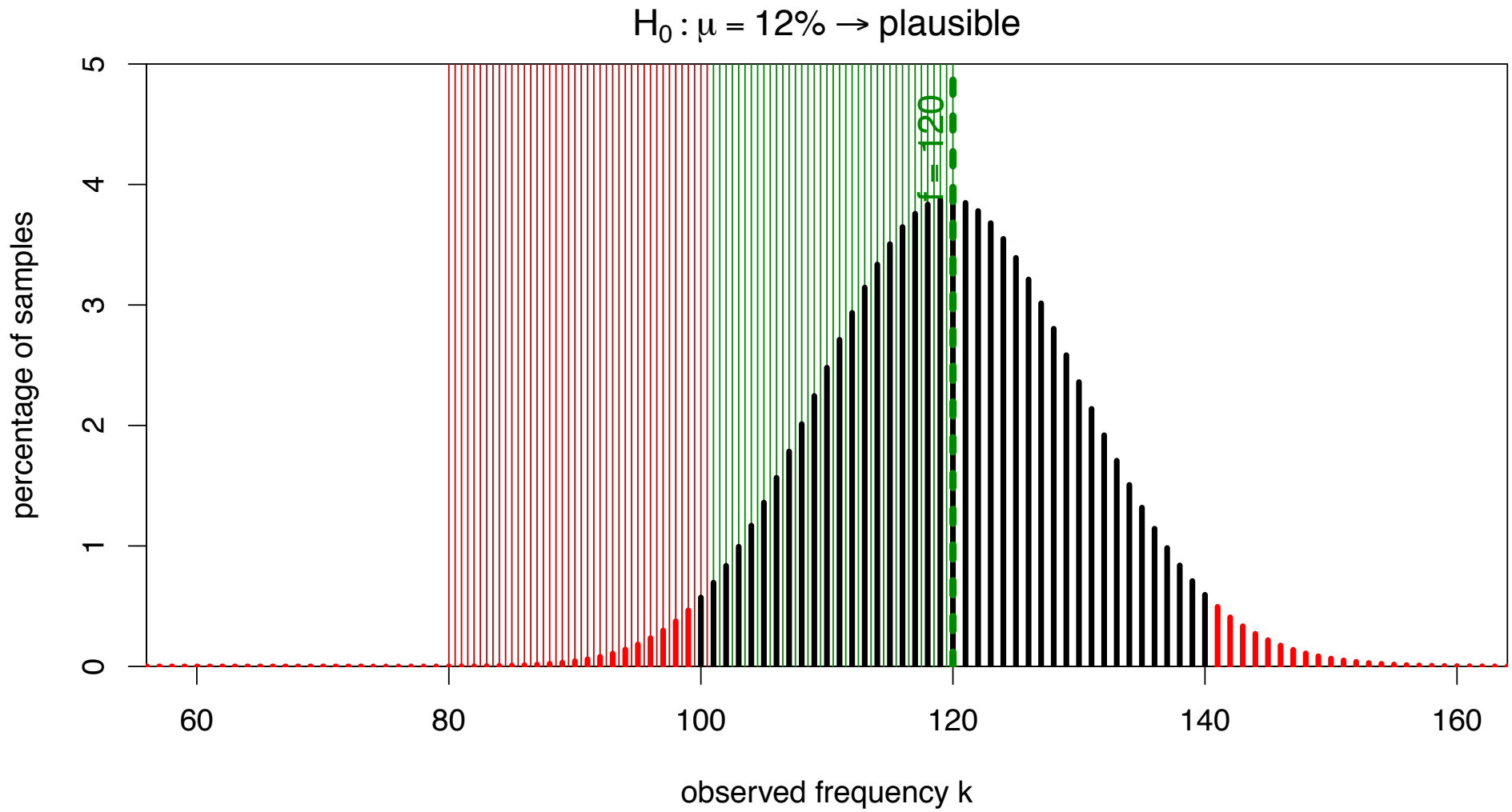
Konfidenz: invertierter Test



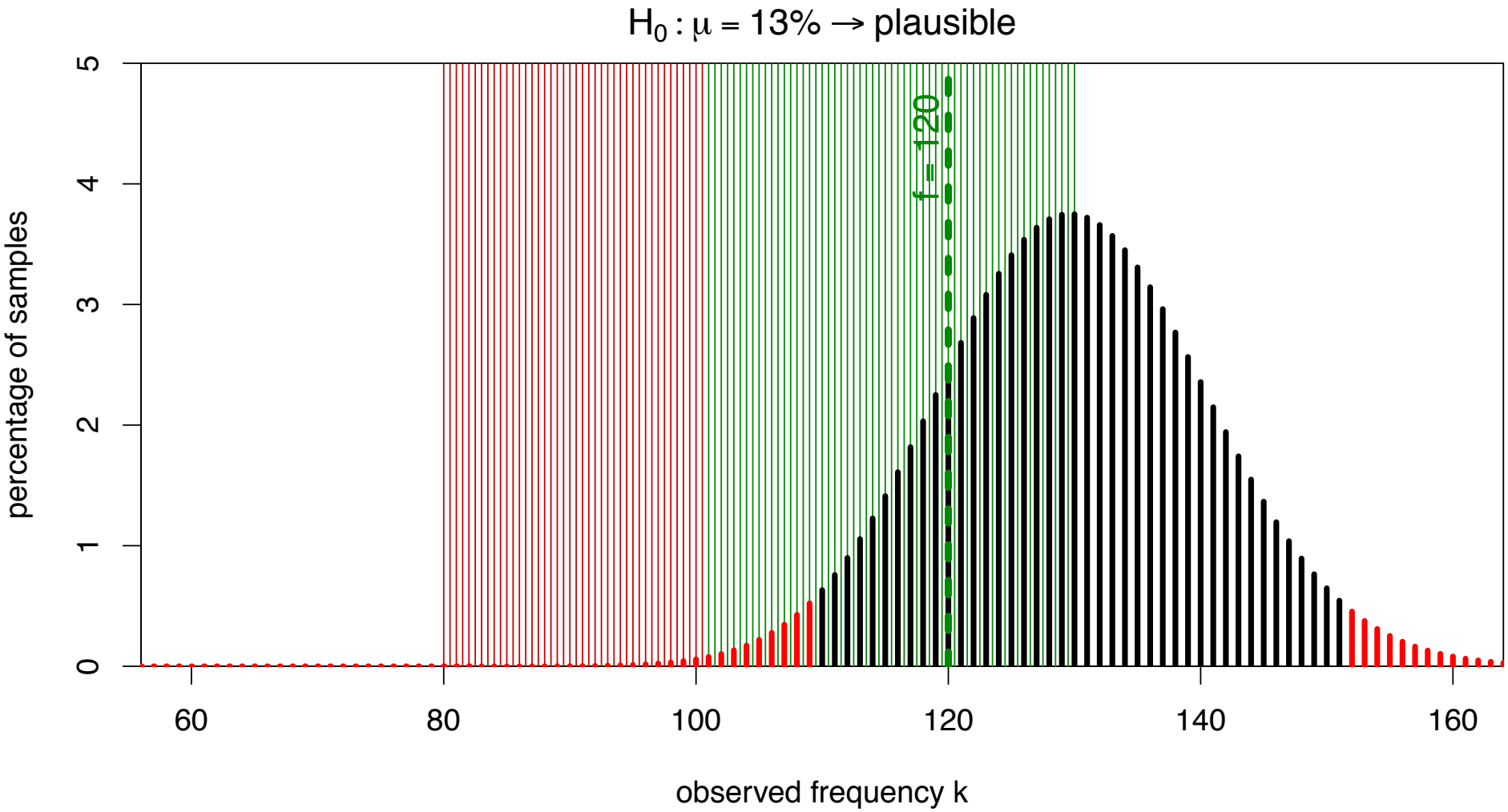
Konfidenz: invertierter Test



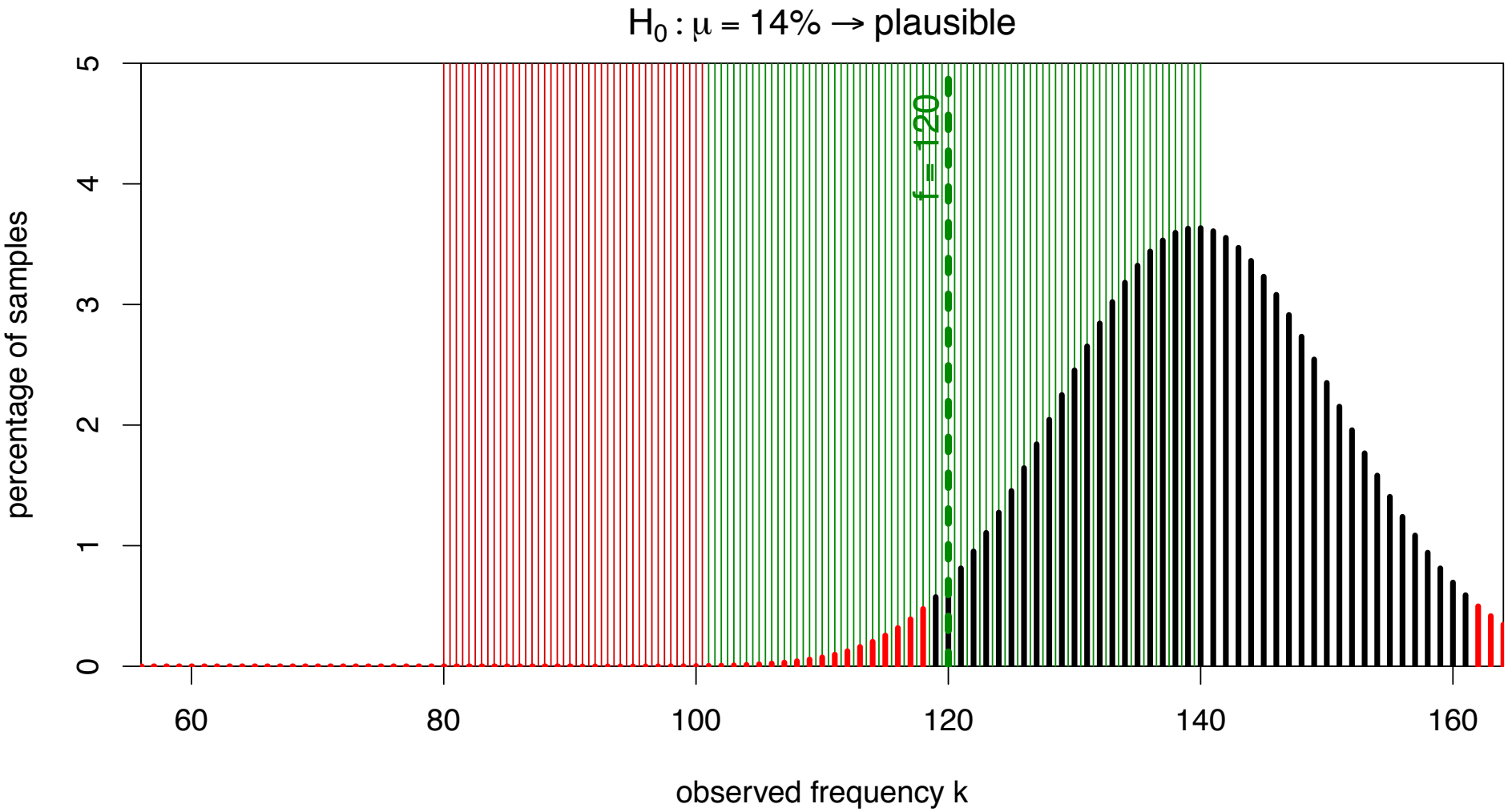
Konfidenz: invertierter Test



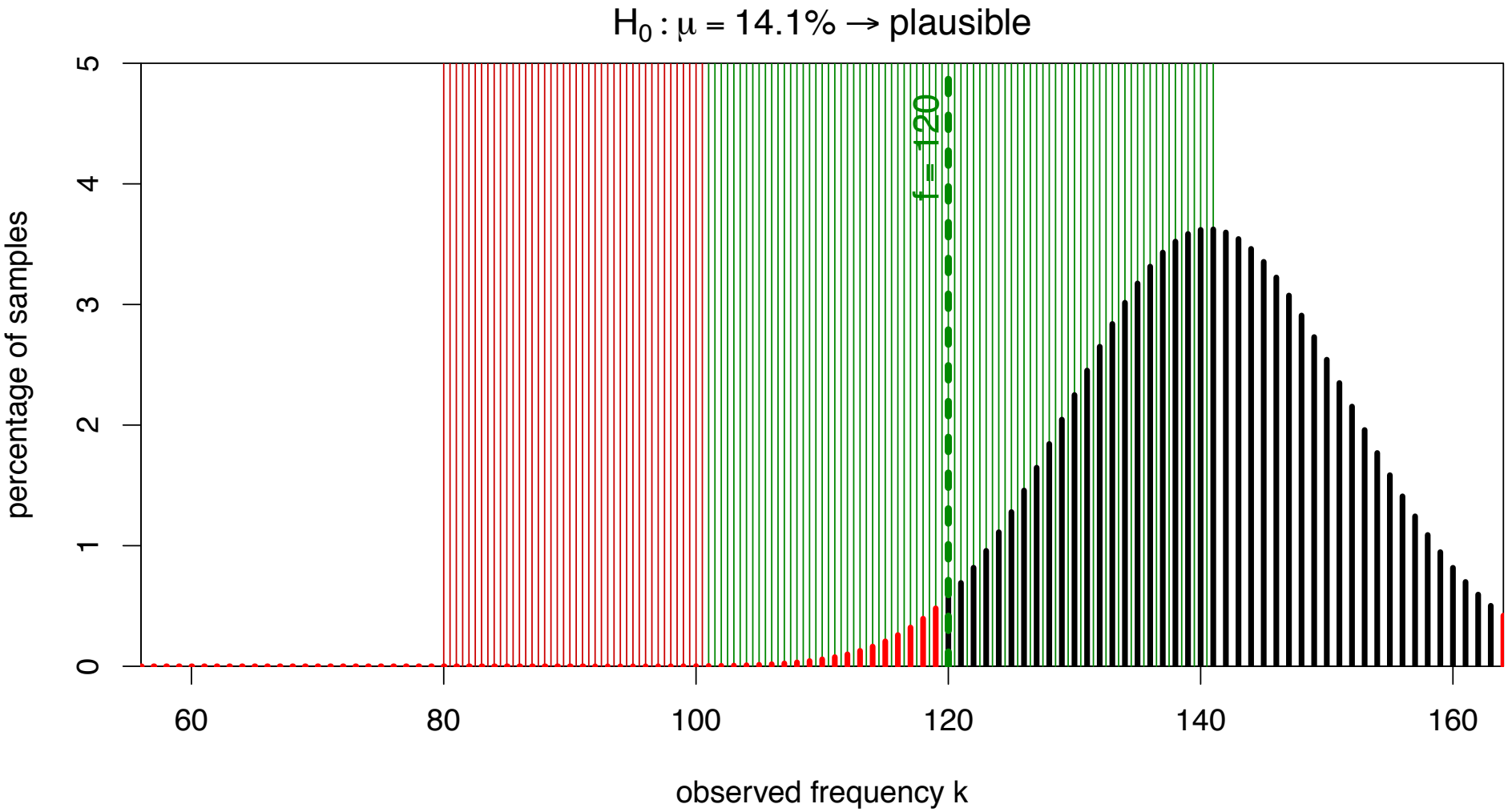
Konfidenz: invertierter Test



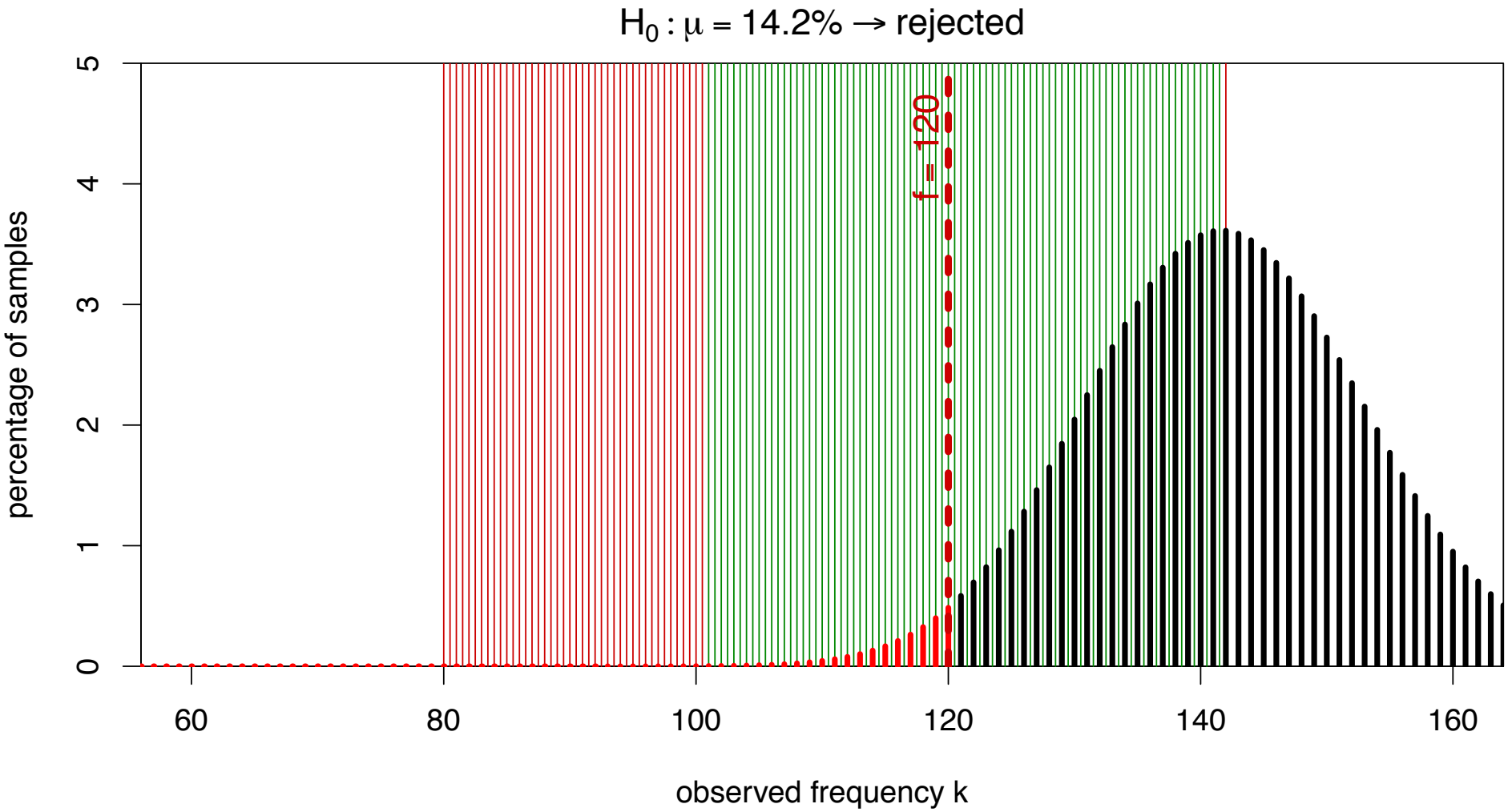
Konfidenz: invertierter Test



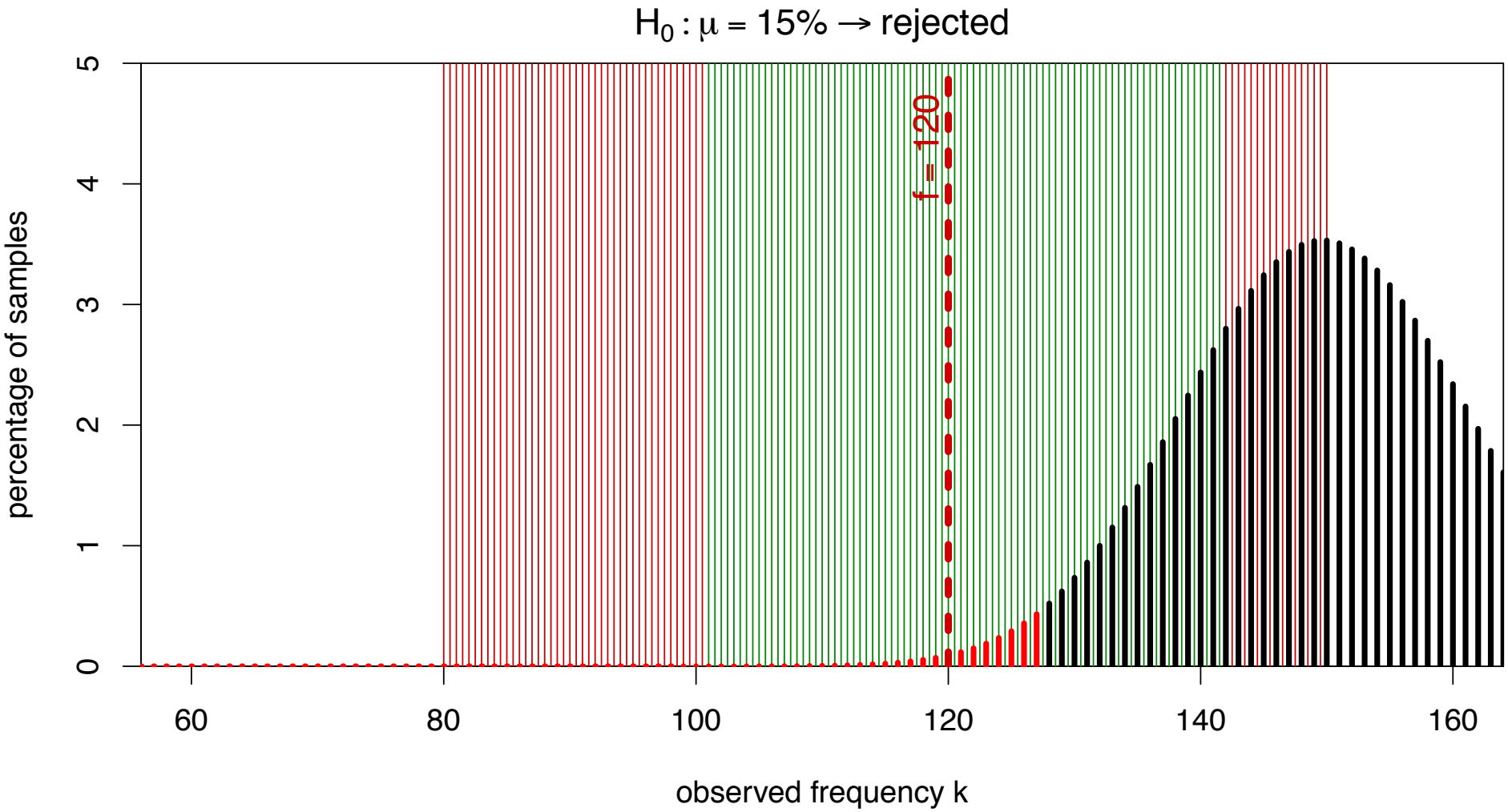
Konfidenz: invertierter Test



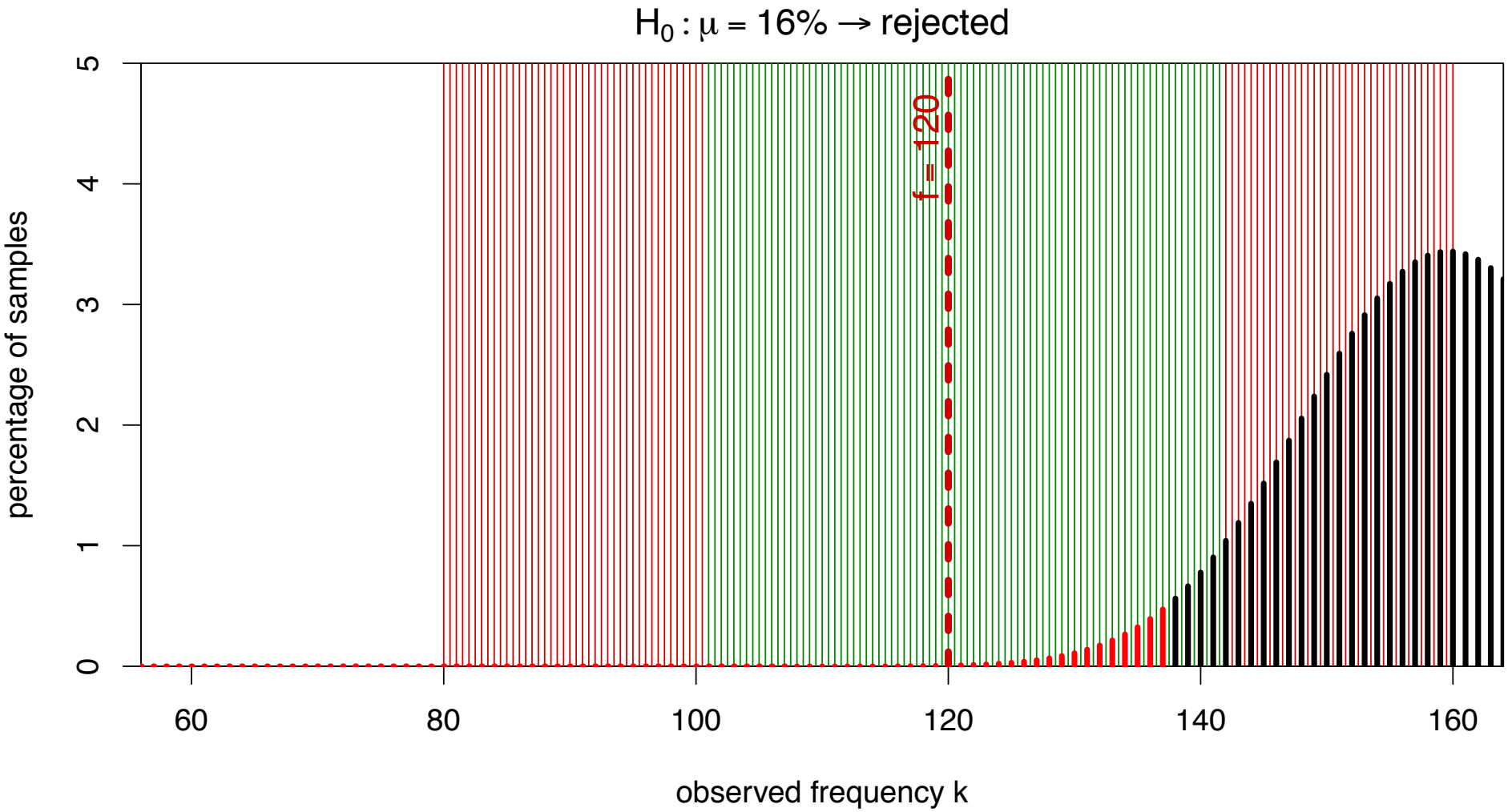
Konfidenz: invertierter Test



Konfidenz: invertierter Test



Konfidenz: invertierter Test



Konfidenzintervall in R

```
> binom.test(120, n = 1000)

    Exact binomial test

data:  120 and 1000
number of successes = 120, number of trials = 1000,
    p-value < 2.2e-16
alternative hypothesis: true probability of success
    is not equal to 0.5
95 percent confidence interval:
 0.1005009 0.1417669
...
```

→ Tatsächliche Häufigkeit im Bereich 10.1% … 14.2%
mit 95% Konfidenz (bei Signifikanzniveau $\alpha = .05$)

Konfidenzintervall für Effektgröße bei Häufigkeitsvergleich

```
> prop.test(c(52, 76), c(500, 500))
    2-sample test for equality of proportions with …

data:  c(52, 76) out of c(500, 500)
X-squared = 4.7395, df = 1, p-value = 0.02948
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.091306444 -0.004693556
sample estimates:
prop 1 prop 2
 0.104  0.152
```

→ Tatsächlicher Unterschied $\mu_2 - \mu_1$ (Effektgröße) ist mind. 0.47%, höchstens 9.1% (Prozentpunkte)

Signfikanz vs. Relevanz

- Erinnerung: Stichprobengröße → Trennschärfe
- **Studie 1:** $k_1 = 42$, $n_1 = 90$ | $k_2 = 76$, $n_2 = 110$
- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$
- Welche Studie ist „interessanter“?

Signfikanz vs. Relevanz

- Erinnerung: Stichprobengröße → Trennschärfe

- **Studie 1:** $k_1 = 42$, $n_1 = 90$ | $k_2 = 76$, $n_2 = 110$

  → p = .002189
- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$

  → p = .000015
- Welche Studie ist „interessanter“?

Signfikanz vs. Relevanz

- Erinnerung: Stichprobengröße → Trennschärfe

- **Studie 1:** $k_1 = 42$, $n_1 = 90$ | $k_2 = 76$, $n_2 = 110$

  → p = .002189 $\quad \mu_2 - \mu_1 \geq 7.97\%$ (Prozentpunkte)

- **Studie 2:** $k_1 = 12500$, $n_1 = 51000$ | $k_2 = 11200$, $n_2 = 48000$

  → p = .000015 $\quad \mu_1 - \mu_2 \geq 0.64\%$ (Prozentpunkte)

- Welche Studie ist „interessanter“?
  - Unterschied bei Studie 2 ist mehr signifikant, aber linguistisch (vermutlich) nicht relevant

ÜBUNGEN

Übung – Komposita in L2-Deutsch

- Wir nehmen nun eine größere Stichprobe, um herauszufinden, ob der Unterschied signifikant ist (Daten aus dem Falko-Korpus, Reznicek et al. 2010)
- Falls kein Zufall dahinter steckt, werden auch weitere Daten einen signifikanten Unterschied aufweisen

Übung

Wir lesen den Datensatz ein:

```
#Daten einlesen
> comp_data <- read.table(file.choose(), header=TRUE, as.is=TRUE,
    fileEncoding="UTF-8")
> head(comp_data)
 ID          token  L1         lemma         head       mod     type
1  1  Geschäftsmann eng Geschäftsmann         Mann Geschäfts compound
2  2      Problemen eng       Problem      Problem      NULL  simplex
3  3        Konzept eng       Konzept      Konzept      NULL  simplex
4  4      Gerichter eng     Gerichter    Gerichter      NULL  unknown
5  5        Gesetze eng        Gesetz       Gesetz      NULL  simplex
6  6 Entscheidungen eng  Entscheidung Entscheidung      NULL  simplex
```

Daten darstellen

Uns interessiert die Verteilung von Komposita/Simplizia:

```
> #Kreuztabelle L1 / Nomentyp erstellen
> L1_type = xtabs(~L1+type, data=comp_data)
> L1_type
 type
L1    compound simplex unknown
  afr       76     806      72
  cat        8      87       4
  ces       39     460      40
  cma       15      84       8
  dan      459    2598     163
  deu     1409    9378     632
  ell       37     557      18
```

Unser 1. Balkendiagramm ☺

- Wir möchten die Proportionen bei jeder L1 betrachten und "unknown" ausblenden

```
> barplot(t(L1_type[,1:2]),cex.names=0.6,las=2)
```

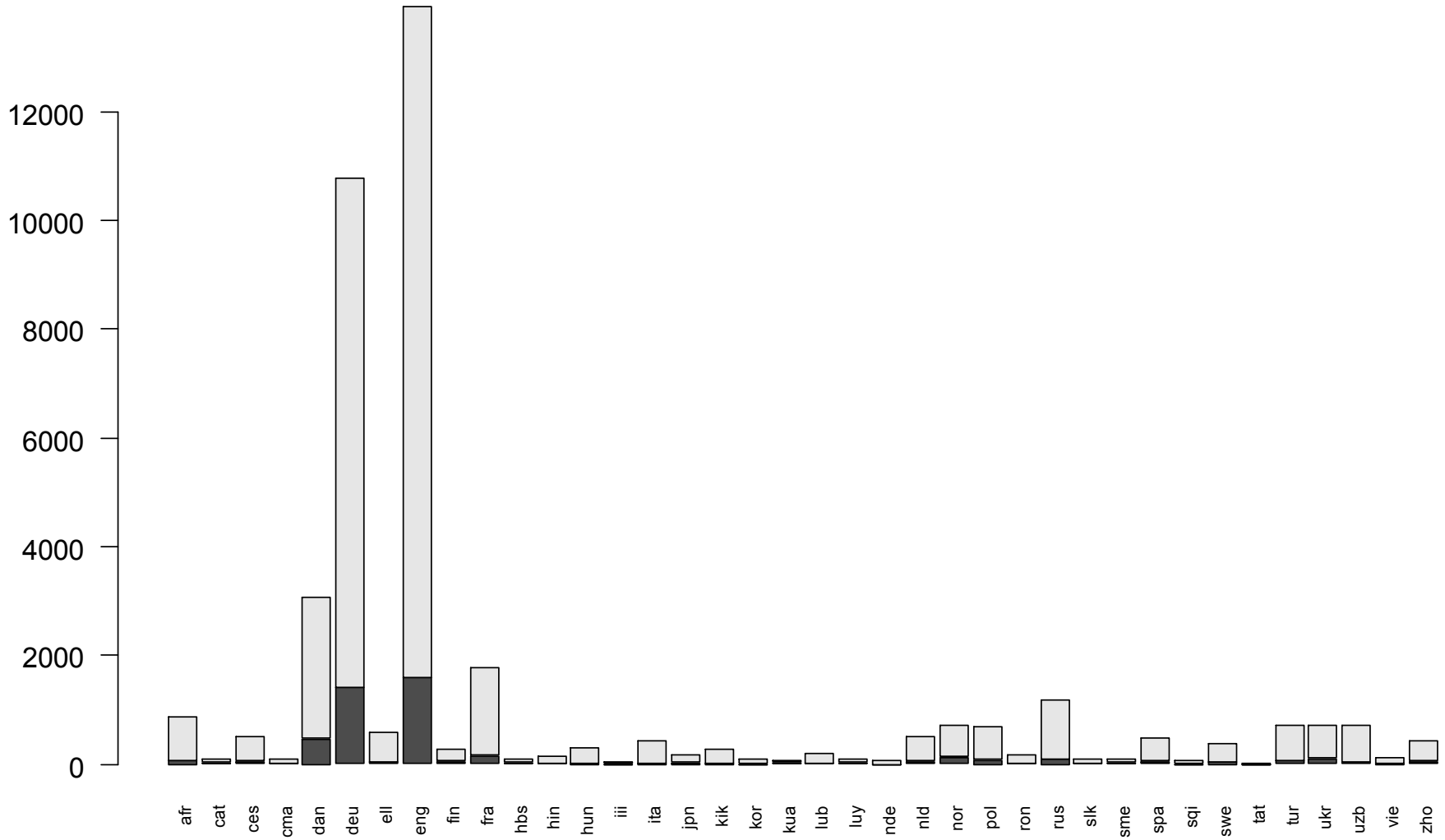
Axenbeschriftung im Lot

Matrix transponieren

Alle Zeilen, Spalten 1 bis 2

Balkennamen verkleinern

Unser 1. Balkendiagramm 😐



Probleme

- Die Darstellung ist unübersichtlich:
  - Balken sehr unterschiedlich groß
  - einige Sprachen haben mehr Daten – kleine Sprachen unsichtbar
  - Einige L1 haben *sehr* wenig Daten
- Lösungen:
  - Daten *normalisieren*
  - Nur einigermaßen gut belegte L1 betrachten

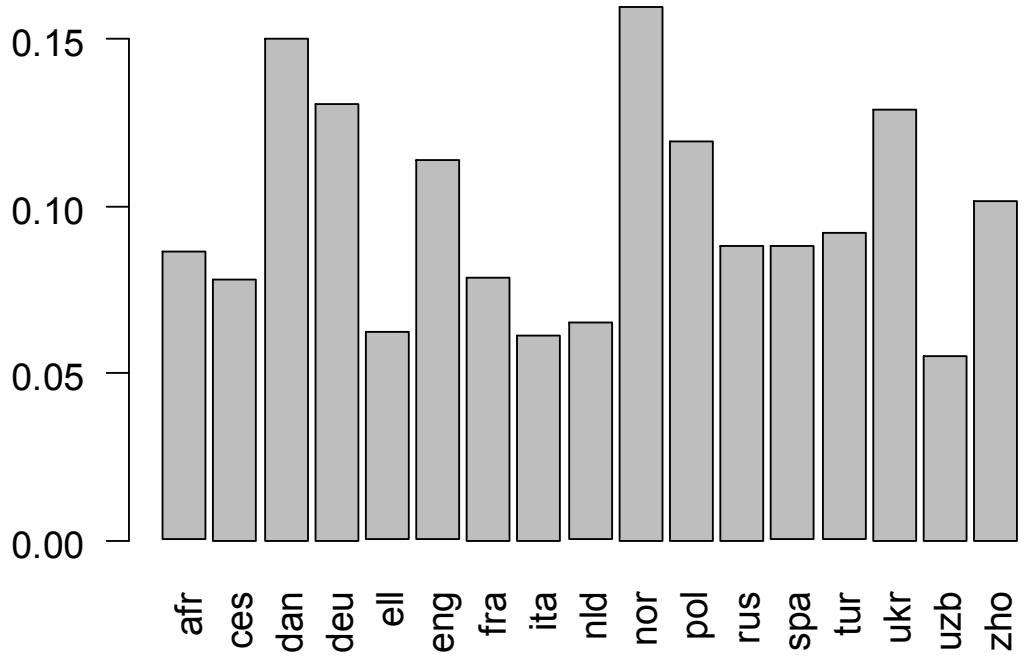
2. Versuch

```
> #Daten filtrieren – mindestens 400 Nomina pro L1
> L1_type2 =   L1_type[L1_type[,1]+L1_type[,
2]>400,1:2]
> L1_type2
      type
L1    compound simplex
  afr       76     806
  ces       39     460
  dan      459    2598
  deu     1409    9378
  ell       37     557
...
```

2. Versuch

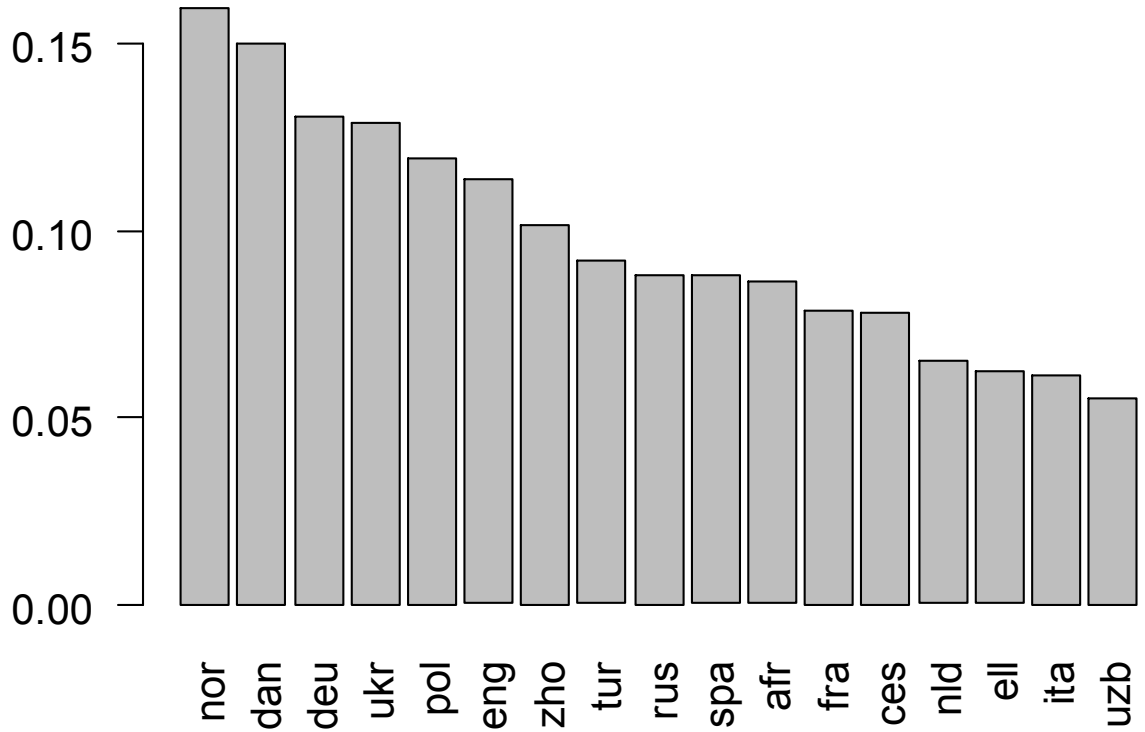
```
#normalisieren – Anzahl Komposita / Nomina insgesamt
> L1_type3 = L1_type2[,1]/(L1_type2[,1]+L1_type2[,2])
> L1_type3
afr         ces         dan         deu     ...
0.08616780 0.07815631 0.15014720 0.13062019

#erneut darstellen
> barplot(L1_type3)
```

Darstellung – unsortiert 😐



Darstellung – sortiert ☺

```
> sorted=sort(L1_type3,decreasing=T)
> barplot(sorted)
```



Verhalten sich die L1-Gruppen gleich?

Hier darf man **keine** normalisierten Daten verwenden (mehr Daten → mehr Sicherheit)

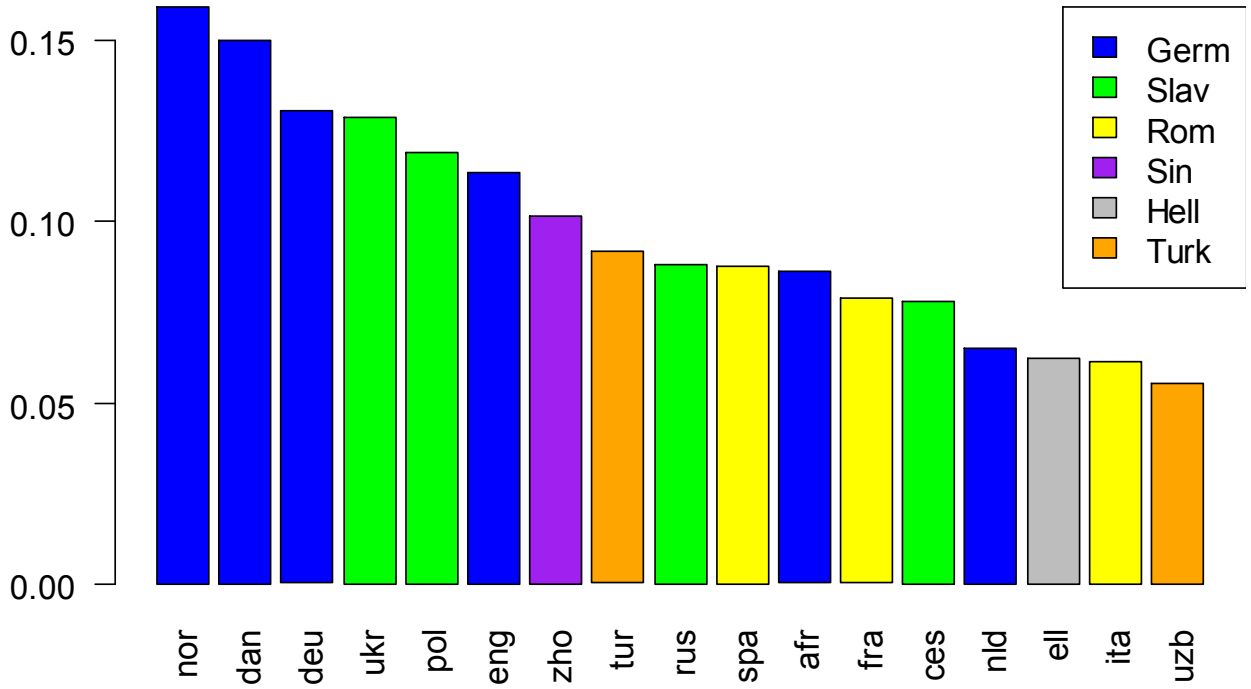
```
> prop.test(L1_type2)

        17-sample test for equality of proportions
   without continuity correction

data:  L1_type2
X-squared = 197.5027, df = 16, p-value < 2.2e-16
alternative hypothesis: two.sided
sample estimates:
    prop 1     prop 2      …
0.08616780 0.07815631 …
```

Zusammenhang mit Sprachfamilie?

- Benutzen bspw. Lerner mit germanischen L1 mehr Komposita?
- Wir machen unser Diagramm bunt, und mit Legende:

```
> barplot(sorted,
   col=c("blue","blue","blue","green","green",
   "blue","purple","orange","green","yellow",
   "blue","yellow","green","blue","grey","yellow",
   "orange"))
> legend("topright",
   c("Germ","Slav","Rom","Sin","Hell","Turk"),
   fill=c("blue","green","yellow","purple","grey",
   "orange"))
```

Sprachfamilien



Verhalten sich Lerner und Muttersprachler gleich?

- Unsere Kreuztabelle hat jede L1 in einer Zeile
- Nun möchten wir eine Kreuztabelle mit nur zwei Zeilen: "deu" oder "nicht deu":

```
> de_type=xtabs(~(L1=="deu")+type,
   data=comp_data)
> de_type
                type
L1 == "deu" compound  simplex   unknown
      FALSE     3284    27023     1743
      TRUE      1409     9378      632
```

prop.test() wiederholen

```
> de_type = de_type[,1:2] #Spalte "unknown" löschen
> prop.test(de_type)

        2-sample test for equality of proportions
    with continuity correction

data:  de_type
X-squared = 38.7566, df = 1, p-value = 4.801e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.02958379 -0.01494098
sample estimates:
   prop 1    prop 2
0.1083578 0.1306202
```

Signifikante Unterschiede

- Mit größeren Stichproben werden selsbt kleine Unterschiede signifikant:
  - Lerner verhalten sich signifikant anders als Muttersprachler
  - Stärke des Unterschieds kann gering sein

Übung

- Unterscheiden sich die Sprachen mit mehr Komposita als "deu" signifikant von den deutschen Daten?

Eine mögliche Lösung

```
> attach(comp_data)
> dan_deu_nor = comp_data[L1=="dan" | L1=="nor" | L1=="deu",]
> ddn_tab=xtabs(~(L1 =="deu")+type, data=dan_deu_nor)[,1:2]
> prop.test(ddn_tab)

        2-sample test for equality of proportions with
continuity correction

data:  ddn_tab
X-squared = 10.5898, df = 1, p-value = 0.001137
alternative hypothesis: two.sided
95 percent confidence interval:
 0.008008357 0.034568865
sample estimates:
   prop 1    prop 2
0.1519088 0.1306202
```

Wortschatz vergleichen

- Benutzen die Lerner weniger **unterschiedliche** Komposita als Muttersprachler?
- **Achtung:**
  - Hier muss man zwei gleich große Stichproben nehmen
  - In größeren Stichproben wird es immer schwieriger, neue Worttypen zu finden
  - Wir nehmen deshalb eine Stichprobe in der Größenordnung der kleinsten zuvergleichenden Gruppe (L1="deu")

Stichprobengröße bestimmen

- Wie viele Daten haben wir für L1="deu"?

```
> length(comp_data[comp_data$L1=="deu" &
comp_data$type=="compound","lemma"])
[1] 1409
```

- Und für die Lerner?

```
> length(comp_data[comp_data$L1!="deu" &
comp_data$type=="compound","lemma"])
[1] 3284
```

Stichproben ziehen

- 1400 L2-Komposita:

```
> L2_comps = sample(comp_data[comp_data$L1!
="deu" & comp_data
$type=="compound","lemma"],1400)
```

- Und 1400 L1-Komposita:

```
> L1_comps = sample(comp_data[comp_data
$L1=="deu" & comp_data
$type=="compound","lemma"],1400)
```

Typen zählen

- Wir machen zwei Tabellen mit den Frequenzen aller Wortformen:

```
> L1_comp_freqs = table(L1_comps)
> L2_comp_freqs = table(L2_comps)
```

- Die Länge von jeder Tabelle entspricht der Typenanzahl: (Zahlen können variieren!)

```
> length(L1_comp_freqs)
[1] 975
> length(L2_comp_freqs)
[1] 824
```

Proportionen gleich?

```
> prop.test(c(824,975),c(1400,1400))

        2-sample test for equality of proportions
with continuity correction

data:  c(824, 975) out of c(1400, 1400)
X-squared = 34.9845, df = 1, p-value = 3.323e-09
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.14384965 -0.07186464
sample estimates:
   prop 1    prop 2
0.5885714 0.6964286
```

Übungsaufgaben

- Wiederholen Sie die Untersuchung mit **allen** Nomina: benutzen Lerner weniger unterschiedliche Nomina insgesamt?
- Wie sieht es aus bei den **Simplizia**?
- Und bei den **Kompositionsköpfen** bzw. -**modifikatoren**?

Endlich!

KAFFEEPAUSE

ASSOZIATION UND UNABHÄNGIGKEITSTEST

Vergleich von Merkmalen

- Bisher: Vergleich von zwei Stichproben
  - aus verschiedenen Grundgesamtheiten
- Jetzt: Vergleich von zwei Merkmalen
  - unterschiedliche Eigenschaften derselben Token
  - eine Stichprobe aus einer Grundgesamtheit
- Fallbeispiel: englische Dativalternation
  - Peter gave [$_{NP}$ his friend] the book
    vs. Peter gave the book [$_{PP}$ to his friend]
  - Besteht ein Zusammenhang mit Informationsstatus?

Dativalternation

- Was für eine Stichprobe wird benötigt?
  - Token = Instanzen von VPen mit Dativobjekt
  - (manuell) annotiert: Dativ-Realisierung (NP/PP), Informationsstatus (new, given, accessible), …
  - aus welchen Textquellen?
- Hier: Teilmenge von Bresnan et al. (2007)
  - *give*-VPen aus *Wall Street Journal* (Zeitungsartikel) und Switchboard-Dialogen (gesprochene Sprache)
  - komplett in R-Paket `languageR` (Baayen 2008)

Dativalternation

```
> Give = read.delim("dative_give.txt")
> Give = read.delim(file.choose()) # alternativ
> dim(Give)
[1] 250   6
> head(Give, 5)
 Recip  AccessRec AccessTheme   AnimRec AnimTheme VerbClass
1     PP        new         new   animate inanimate  transfer
2     PP      given  accessible   animate inanimate  transfer
3     NP        new  accessible inanimate inanimate  abstract
4     PP        new         new inanimate inanimate  abstract
5     NP accessible  accessible   animate inanimate  abstract
```

Dativalternation

```
> summary(Give) # Überblick über Kategorien

 Recip          AccessRec       AccessTheme
 NP:199   accessible: 70   accessible:134
 PP: 51   given     :146   given     : 15
          new       : 34   new       :101

     AnimRec         AnimTheme           VerbClass
 animate  :209   animate  :  2   abstract     :224
 inanimate: 41   inanimate:248   communication: 10
                                 transfer     : 16
```

Assoziation

- Wie können wir feststellen, ob es einen Zusammenhang zwischen den Merkmalen Recip und AccessRec gibt? (Assoziation)
- Wie oft kommen bestimmte Merkmals-ausprägungen miteinander vor?
  - sog. Kreuztabelle („contingency table“)
  - Wie erstellt man eine Kreuztabelle in R?

```
> kt = xtabs(~ Recip + AccessRec, data=Give)
```

Kreuztabellen

```
> kt = xtabs(~ Recip + AccessRec, data=Give)
> print(kt)
```

	access	given	new
NP	48	137	14
PP	22	9	20

- Zusammenhang erkennbar?
- Woran?
- Insgesamt oder für einzelne Felder?
- Signifikanz?

Statistische Unabhängigkeit

- Ein Zusammenhang besteht, wenn bestimmte Merkmalskombinationen auffallend häufig oder selten auftauchen
  - sog. Kookkurrenz von Merkmalsausprägungen
- Hängt von Häufigkeit einzelner Kategorien ab
  - Erwartung: viele Kookkurrenzen bei zwei häufigen Kategorien, wenige bei zwei seltenen Kategorien
  - Annahme: Merkmale sind statistisch unabhängig
  - mathematische Formel für erwartete Häufigkeit

Notation für Kreuztabellen

```
> kt = xtabs(~ Recip + AccessRec, data=Give)
> print(kt)
```

	access	given	new
NP	48	137	14
PP	22	9	20

	access	given	new	
NP	n_{11}	n_{12}	n_{13}	$= n_{1\bullet}$
PP	n_{21}	n_{22}	n_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Notation für Kreuztabellen

```
> kt = xtabs(~ Recip + AccessRec, data=Give)
> print(kt)
> addmargins(kt)
```

	access	given	new	
NP	48	137	14	= 199
PP	22	9	20	= 51
	= 70	= 146	= 34	250

	access	given	new	
NP	n_{11}	n_{12}	n_{13}	$= n_{1\bullet}$
PP	n_{21}	n_{22}	n_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Erwartete Häufigkeit

- H_0: Unabhängigkeitshypothese
  - Wk für NP = $n_{1\bullet} / n$
  - Wk für given = $n_{\bullet 2} / n$
  - Kookkurrenz-Wk = $n_{1\bullet}\, n_{\bullet 2} / n^2$
  - Erwartete Häufigkeit e_{12} für n Token

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

E	access	given	new	
NP	e_{11}	e_{12}	e_{13}	$= n_{1\bullet}$
PP	e_{21}	e_{22}	e_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Erwartete Häufigkeit

- Erwartete Häufigkeit in R (für Profis)

```
> n_rows = rowSums(kt)
> n_cols = colSums(kt)
> n = sum(kt)
> kt.e = outer(n_rows, n_cols) / n
> round(kt.e, 1)
> addmargins(kt.e)
```

$$e_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

E	access	given	new	
NP	e_{11}	e_{12}	e_{13}	$= n_{1\bullet}$
PP	e_{21}	e_{22}	e_{23}	$= n_{2\bullet}$
	$= n_{\bullet 1}$	$= n_{\bullet 2}$	$= n_{\bullet 3}$	n

Assoziation & Chi-Quadrat-Test

- Vergleich von erwarteten und tatsächlichen Häufigkeiten
- Hypothesentest für H_0: Merkmale sind unabhängig
- Chi-Quadrat-Statistik → Signifikanzwert

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

E	access	given	new	
NP	55.7	116.2	27.1	= 199
PP	14.3	29.8	6.9	= 51
	= 70	= 146	= 34	250

	access	given	new	
NP	48	137	14	= 199
PP	22	9	20	= 51
	= 70	= 146	= 34	250

Chi-Quadrat-Test in R

```
> X2 = sum( (kt - kt.e)^2 / kt.e )
> print(X2)
> ergebnis = chisq.test(kt)
> print(ergebnis)
```

$$X^2 = \sum_{ij} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

```
   Pearson's Chi-squared test

data:  kt
X-squared = 54.376, df = 2, p-value = 1.557e-12
```

Resultat: $p = 1.557 \times 10^{-12} = .000000000001557$

Chi-Quadrat-Test in R

- Welche Merkmalskombinationen sind auffällig?
- Standardisierte Abweichung (z-score) für jede Kombination
- Unter H_0: jedes z_{ij} folgt einer Standardnormalverteilung

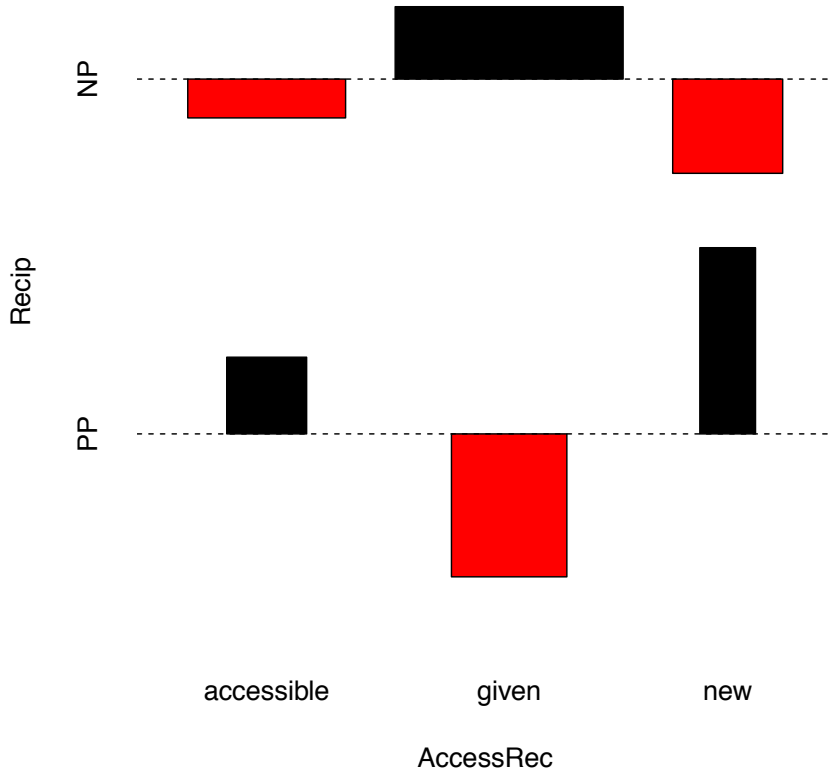
$$z_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

```
> round(ergebnis$expected, 1)
> round(ergebnis$residuals, 2) # = z_ij
```

Chi-Quadrat-Test in R

```
> round(ergebnis$residuals, 2)
> 2 * pnorm(2.51, lower=FALSE) # für NP / new
> assocplot(t(kt)) # t() = transponieren
```

Z	access	given	new
NP	–1.03	1.93	–2.51
PP	2.04	–3.81	4.96



Signifikanz vs. Relevanz

- Auch standardisierte Abweichung z_{ij} ist ein Signifikanzwert → Trennschärfe
  - Wie kann man die Effektgröße messen?
- Globales Maß: *V*-Koeffizient von Cramér

```
> X2 = ergebnis$statistic
> n = sum(kt)
> faktor = min(dim(kt)) - 1
> V = sqrt(X2 / (n * faktor))
> V  # Wertebereich 0 … 1
```

$$V = \sqrt{\frac{X^2}{n \cdot \min\{r-1, c-1\}}}$$

Mini-Übung

- Gibt es Evidenz für andere Einflüsse auf die Dativalternation (z.B. Belebtheit)?
- Zusammenhang zw. Informationsstatus von Dativobjekt und Akkusativobjekt?
- Ändert sich der Signifikanzwert bei einer größeren Stichprobe? Und Cramérs *V*?

Mini-Übung: Lösungen

```
> kt2 = xtabs(~ Recip + VerbClass, data=Give)
> chisq.test(kt2)

        Pearson's Chi-squared test

data:  kt2
X-squared = 3.8243, df = 2, p-value = 0.1478 (?)

Warning message:
In chisq.test(kt2) : Chi-squared approximation may be
incorrect
> fisher.test(kt2) # exakter Test (aufwendig)

        Fisher's Exact Test for Count Data

data:  kt2
p-value = 0.1301
alternative hypothesis: two.sided
```

Mini-Übung: Lösungen

```
> kt3 = xtabs(~ AccessRec+AccessTheme, data=Give)
> print(kt3)
 AccessTheme
AccessRec    accessible given new
  accessible         49     2  19
  given              76    10  60
  new                 9     3  22
> chisq.test(kt3)
    Pearson's Chi-squared test
data:  kt2
X-squared = 18.0605, df = 4, p-value = 0.001201
Warning message: …
> fisher.test(kt3)
```

Mini-Übung: Lösungen

```
> cramer = function (kt) {
    X2 = chisq.test(kt)$statistic
    n = sum(kt)
    faktor = min(dim(kt)) - 1
    sqrt(X2 / (n * faktor))
  }

> chisq.test(kt2)$p.value
[1] 0.001200939
> cramer(kt2)
0.1900553
> chisq.test(10 * kt2)$p.value
[1] 5.528046e-38
> cramer(10 * kt2)
0.1900553
```

Jetzt ernsthaft:

ÜBUNGEN

Mehrere kategoriale Variablen

- Mit prop.test() haben wir nur Gruppen im Hinblick auf einen Wert verglichen
- Wir hatten keinen Weg, mehr als zwei Variablenausprägungen zu behandeln
  (Kompositum oder nicht; was ist mit Simplex, Derivation und Komposition? Zwei- bzw. mehrgliedrige Komposita?)
- Was machen wir, wenn wir mehr haben?

Übung

- Wir arbeiten als nächstes mit informationsstrukturell annotierten Daten
- Zwei Variablen: (Guidelines des SFB632, Dipper et al. 2007)
  - infstat: **giv**(en), **new**, **acc**(essible), **idiom**
  - topic: **ab**(outness), **fs** = framesetter, **nt** = not-topic
- Genauere Unterteilung in "active" bzw. "inactive", "inferrable", "generic" etc.

Übung

- Forschungsfragen:
  - Wie hängt Topikalität mit Bekanntheit zusammen?
  - Gibt es Unterschiede dabei zwischen Vorerwähntheit und Erschließbarkeit?
  - Framesetter und Aboutness-Topik?
- Formulierung von Hypothesen
- Erstellung von passenden Kreuztabellen

Die Daten

```
> infstat_data = read.table(file.choose(), header=TRUE, fileEncoding="UTF-8", as.is=TRUE)
> attach(infstat_data)
> head(infstat_data)
 ID                   referent infstat topic infstat_fine
1  1           Die_Jugendlichen     new    ab          new
2  2                     Zossen     new    ab          new
3  3              ein_Musikcafé     new    nt          new
4  4                        Das     giv    nt   giv-active
5  5                        sie     giv    ab   giv-active
6  6 der_ersten_Zossener_Runde     new    fs          new
```

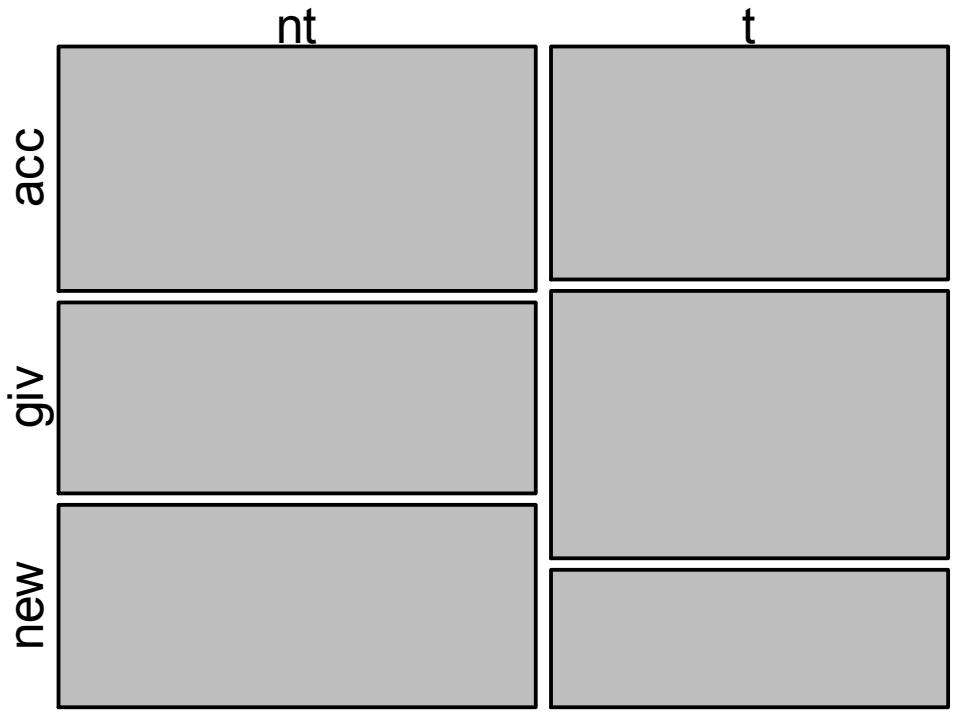
Erste Hypothese

- Frage: hängt Informationsstatus mit Topikalitätzusammen?
- Nullhypothese **H0**: Kein Zusammenhang zwischen den Variablen
- Wir fangen an mit einer groben Untersuchung: binäre Topikalität und Neuheit

Erste Tabelle

```
> no_idiom = subset(infstat_data,infstat!="idiom")
> topic_infstat = xtabs(~ ifelse(topic=="nt","nt","t") + infstat, data = no_idiom)
> topic_infstat

    acc giv new
  nt  72  56  60
  t   57  65  34
```

plot(topic_infstat)



Was ist die Nullhypothese?

- Heißt kein Zusammenhang: alles in jeder Zelle der Tabelle gleich?
- Wir fügen Zwischensummen hinzu:

```
> addmargins(topic_infstat)

       acc giv new Sum
  nt    72  56  60 188
  t     57  65  34 156
  Sum 129 121  94 344
```

- Erwarten wir 344/6 = 57.33 in jeder Kombination?

Was ist die Nullhypothese?

- Nicht wirklich: es kann sein, dass es mehr Nicht-Topiks gibt als Topiks
- Es kann sein, dass die meisten Referenten neu sind
- Aber es darf keine Interaktion geben (mehr neu wenn nicht Topik)

```
> chisq.test(topic_infstat)$expected
```

```
       acc      giv      new
nt 70.5 66.12791 51.37209
t  58.5 54.87209 42.62791
```

gleich verteilt

gleich verteilt

Zusammenhang testen

```
> chisq.test(topic_infstat)

        Pearson's Chi-squared test

data:  topic_infstat
X-squared = 6.6862, df = 2, p-value = 0.03533
```

- Irgendwo sind Zeilen und Spalten **nicht** unabhängig, H0 gilt nicht

Residuen

- Die residuen drücken den Unterschied zwischen Erwartung und Beobachtung aus:
  - Residuals = (observed - expected) / sqrt(expected)

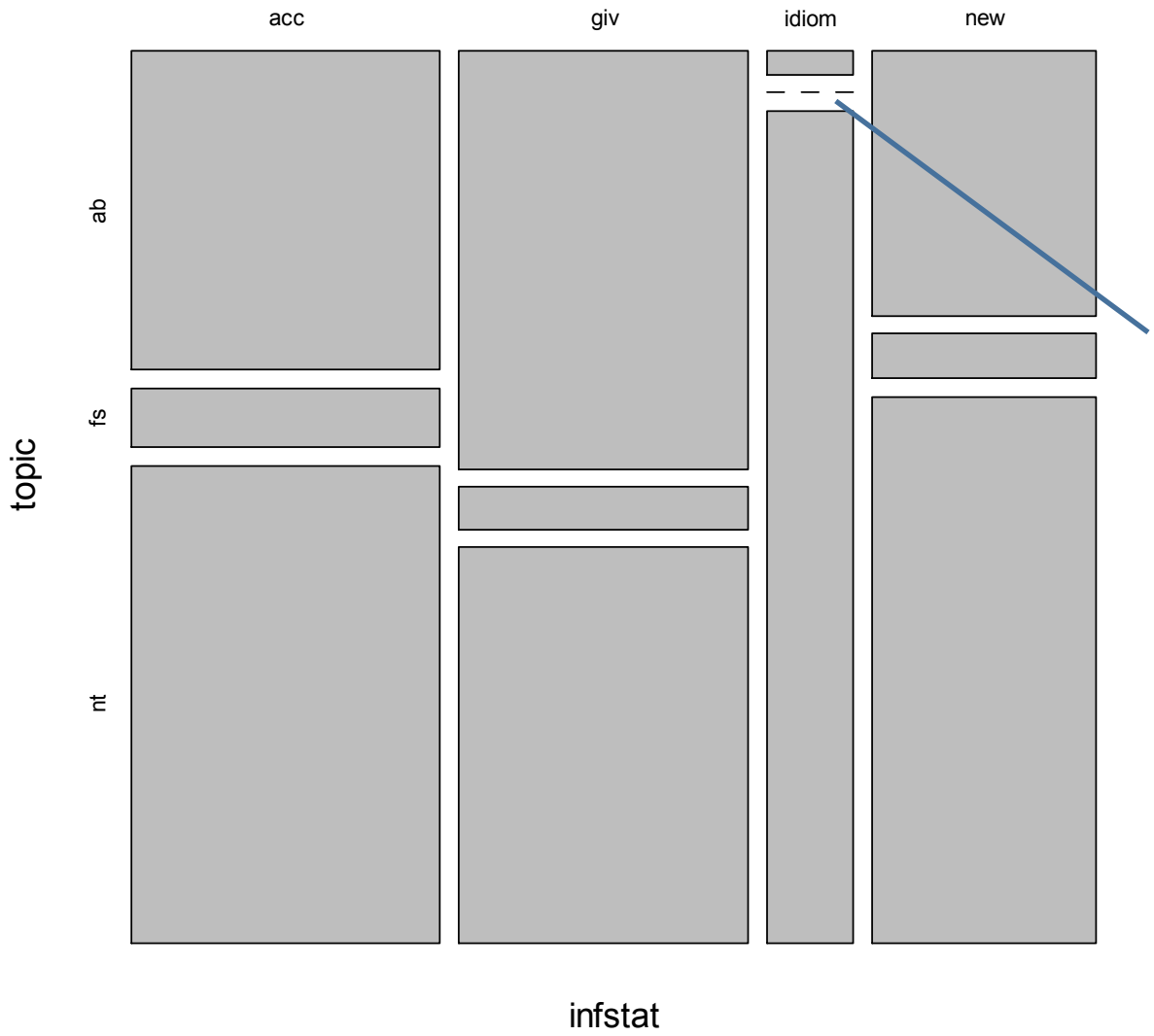
```
> chisq.test(topic_infstat)$residuals

            acc        giv        new
nt   0.1786474 -1.2454529  1.2037653
t   -0.1961161  1.3672374 -1.3214735
```

Zweite Kreuztabelle

- Bekommen wir ein besseres Bild mit allen Kategorien von topic ~ infstat?

```
> is_tab = xtabs(~ infstat + topic, data =
infstat_data)
> is_tab
        topic
infstat ab fs nt
  acc   48  9 72
  giv   59  6 56
  idiom  1  0 35
  new   29  5 60

> plot(is_tab)
```

plot(is_tab)


acc
giv
idiom
new
ab
fs
nt
topic
infstat
Keine Fälle von idiom & fs

Wann gibt es "ab" und "idiom"?

```
> infstat_data[infstat_data$infstat=="idiom" &
infstat_data$topic=="ab",]
     ID     referent infstat topic infstat_fine
155 155 Das_Füllhorn   idiom    ab        idiom
```

- Satz:

"***Das Füllhorn*** *schließlich schüttet man über Kitas und Schulen aus.*"

chisq.test

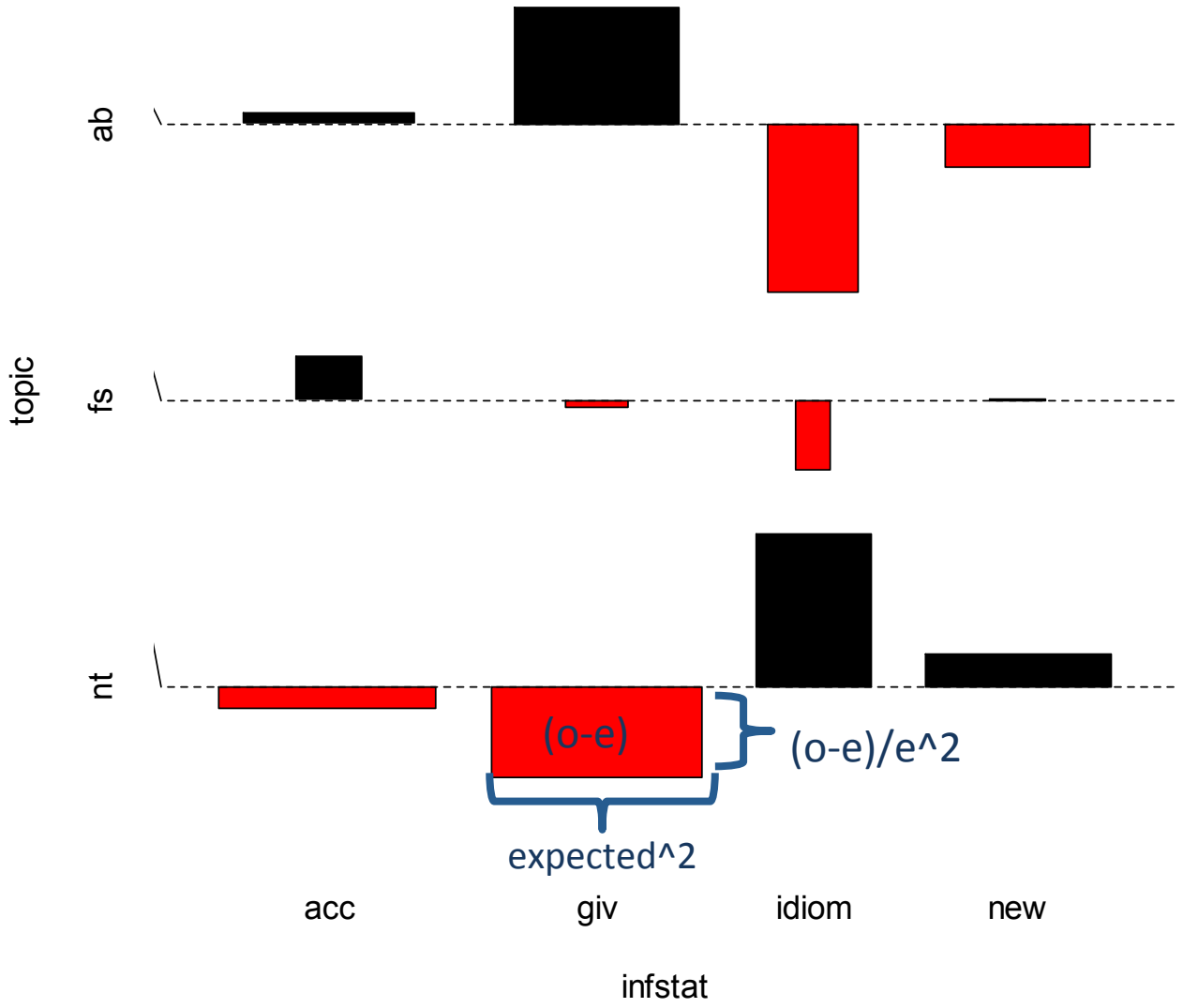
- Gibt es Zusammenhänge mit allen Variablenausprägungen?

```
> is_test = chisq.test(is_tab)
Warning message:
In chisq.test(is_tab) : Chi-squared approximation may be incorrect

> is_test

        Pearson's Chi-squared test

data:  is_tab
X-squared = 32.7535, df = 6, p-value = 1.17e-05
```

Welche Zusammenhänge fallen auf?

```
> is_test$residuals
        topic
infstat           ab          fs           nt
  acc     0.21879436  0.84835502 -0.42555433
  giv     2.32804365 -0.14599183 -1.78101040
  idiom  -3.32505567 -1.37649440  3.01842175
  new    -0.83990410  0.02366243  0.65123443
```

assocplot(is_tab)



Warnung?

- Zu wenig Daten zu den Idiomen:

```
> is_tab
        topic
infstat ab fs nt
   acc   48  9 72
   giv   59  6 56
   idiom  1  0 35
   new   29  5 60
```

Genauerer Test mit fisher.test()

```
> fisher.test(xtabs(~ topic+infstat, data=infstat_data), workspace=2e6)

        Fisher's Exact Test for Count Data

data:  xtabs(~topic + infstat, data = infstat_data)
p-value = 9.734e-07
alternative hypothesis: two.sided
```

Weitere Übungen

- Was passiert, wenn man die noch feineren Kategorien in *infstat_data$infstat_fine* nimmt?
- Was passiert, wenn man acc und giv zusammen mit new kontrastiert?

SCHLUSSWORTE + EVALUATION

Literaturempfehlungen

Für den Einstieg:

- Gries, Stefan Th. (2008). *Statistik für Sprachwissenschaftler.* Göttingen: Vandenhoeck & Ruprecht.
- Gries, Stefan Th. (2009). *Statistics for Linguistics with R: A Practical Introduction.* Berlin: Mouton de Gruyter, Berlin.
- Oakes, Michael P. (1998). *Statistics for Corpus Linguistics.* Edinburgh: Edinburgh University Press.
- Bortz, Jürgen (2005). *Statistik für Sozialwissenschaftler.* Heidelberg: Springer.

Fortgeschrittene Methoden:

- Baayen, R. Harald (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics.* Cambridge: Cambridge University Press.
- Rietveld, Toni & van Hout, Roeland (2005). *Statistics in Language Research: Analysis of Variance.* Berlin/New York: Mouton de Gruyter.

Online-Materialien

- Handouts & Daten zum Tutorium:
  http://wordspace.collocations.de/doku.php/corpus_tutorial:dgfs2012
- SIGIL-Kurs (Baroni & Evert):
  http://sigil.r-forge.r-project.org/
- Butler, C. (1985). *Statistics in Linguistics.* Oxford: Blackwell.
  http://www.uwe.ac.uk/hlss/llas/statistics-in-linguistics/bkindex.shtml
- Galtonbrett-Simulation:
  http://www.math.psu.edu/dlittle/java/probability/plinko/
- R-Homepage (u.a. Lehrbücher & Online-Tutorien):
  http://www.r-project.org/